David J. Klotzkin

Introduction to Semiconductor Lasers for Optical Communications

An Applied Approach

Second Edition

Springer

Introduction to Semiconductor Lasers for Optical Communications

David J. Klotzkin

Introduction to Semiconductor Lasers for Optical Communications

An Applied Approach

Second Edition

Springer

David J. Klotzkin
Department of Electrical
and Computer Engineering
Binghamton University
Binghamton, NY, USA

ISBN 978-3-030-24500-9 ISBN 978-3-030-24501-6 (eBook)
https://doi.org/10.1007/978-3-030-24501-6

1st edition: © Springer Science+Business Media New York 2014
2nd edition: © Springer Nature Switzerland AG 2020
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Nobody questions the importance of semiconductor lasers. The information they transmit is the backbone of the World Wide Web, and they are increasingly finding new applications in solid-state lighting and in spectroscopy, and at new wavelengths ranging all the way from the ultraviolet on gallium nitride to the extremely long wavelengths produced by quantum cascade lasers. Even in optical communications, lasers are used in different ways, from metropolitan links using directly modulated devices to Tb/s transmission systems incorporating advanced detection and modulation schemes.

In this book, I introduce semiconductor lasers from an operational perspective to those who have a background in engineering or optics, but no familiarity with lasers. The objective here is to present semiconductor lasers in a way that are both accessible and interesting to advanced undergraduate and graduate students. The target audience for this book is someone who is potentially interested in careers in semiconductor lasers, and the decision of what topic to cover is driven both by the importance of the topic and how fundamental it is to the whole field. I hope to make the reader very comfortable with both the scientific and engineering aspects of this discipline.

The topics and emphasis were selected based largely on my experience in the semiconductor laser industry. My goal is that after reading the book, the reader appreciates most of the aspects of laser fabrication and performance and can get immediately and actively involved in the engineering of these devices.

The book starts with talking generally about optical communications and the need for semiconductor lasers. It then discusses the general physics of lasers and moves on to the relevant specifics of semiconductors. There are chapters on optical cavities, direct modulation, distributed feedback, and electrical properties of semiconductor lasers. Topics like fabrication and reliability are also covered. This second edition also includes a discussion of optical communication, including amplitude-modulated and coherent formats.

The book is appropriate as the primary text for a one-semester course on semiconductor lasers at the advanced undergraduate or introductory graduate level, or would also be appropriate as one of the texts in a general course in photonics, optoelectronics, or optical communications.

Despite all care, errors have a way of creeping in. I apologize in advance for any errors which may remain. Should any error be discovered, readers are invited to bring it to my attention and I will maintain a list of errata for the benefit of those using the book or a possible future edition.

Binghamton, USA David J. Klotzkin

Acknowledgements

First, much thanks to Springer, and Michael Luby, for the opportunity to do a second edition of this book.

My background in lasers is largely from commercial laser communication companies. I appreciate the opportunities I have had to work at Lasertron (later acquired by Corning), Lucent (which later became Agere), Ortel (which later became part of Agere, and then part of Emcore), Binoptics (which was acquired by Macom), Finisar (currently being acquired by II-VI), and Source Photonics. Change is constant in the laser industry. At all of these places, there were always laser problems to work on and people to learn from! In my commercial career I had the chance to work with many knowledgeable and helpful people, particularly Malcolm Green, Phil Kiely, Hanh Lu, Julie Eng, Richard Sahara, Jia-Sheng Huang, Tsurugi Sudo, Yashiro Matsui, John Bai, Martin Kwakernaak, and Ashish Verma. A particular thanks to Binoptics and Finisar for allowing me to use some data in this book.

I very much appreciate Sylvia Smolorz generously sharing her time and expertise on some topics.

I am happy to again thank Mary Lanzerotti for her help at all phases of this project, from the start and through the first and then this second edition. Her input has very much improved the book. A thank you also to James Pitarresi and Stephen Cain for some of the pictures.

My laser course and students were always the motivation for this work. I appreciate what I have heard from my students in the course over the years and hope the presentation is as clear as they deserve.

Much thanks to my family, for their support over the time this has taken.

Let me finally thank my graduate research advisor, Prof. Pallab Bhattacharya, for getting me started on this fascinating field.

Contents

1 Introduction: The Basics of Optical Communications 1
1.1 Introduction 1
1.2 Introduction to Optical Communications 2
1.2.1 The Basics of Optical Communications 2
1.2.2 A Remarkable Coincidence 4
1.2.3 Optical Amplifiers 5
1.2.4 A Complete Technology 6
1.3 A Picture of Semiconductor Lasers 6
1.4 Organization of the Book 8
1.5 Summary and Learning Points 9
1.6 Questions and Problems 10

2 The Basics of Lasers 11
2.1 Introduction 11
2.2 Introduction to Lasers 11
2.2.1 Black Body Radiation 12
2.2.2 Statistical Thermodynamics Viewpoint of Black Body Radiation 13
2.2.3 Some Probability Distribution Functions 14
2.2.4 Density of States 15
2.2.5 Spectrum of a Black Body 19
2.3 Black Body Radiation: Einstein's View 19
2.4 Implications for Lasing 21
2.5 Differences Between Spontaneous Emission, Stimulated Emission, and Lasing 24
2.6 Some Example of Laser Systems 25
2.6.1 Erbium-Doped Fiber Laser 25
2.6.2 HeNe Gas Laser 26
2.7 Summary and Learning Points 28
2.8 Questions 29
2.9 Problems 29

3 Semiconductors as Laser Materials 1: Fundamentals 31
3.1 Introduction 31
3.2 Energy Bands and Radiative Recombination 32
3.3 Semiconductor Laser Material System 34
3.4 Determining the Band Gap 36
3.4.1 Vegard's Law: Ternary Compounds 37
3.4.2 Vegard's Law: Quaternary Compounds 39
3.5 Lattice Constant, Strain, and Critical Thickness 39
3.5.1 Thin Film Epitaxial Growth 40
3.5.2 Strain and Critical Thickness 41
3.6 Direct and Indirect Bandgaps 43
3.6.1 Dispersion Diagrams 44
3.6.2 Features of Dispersion Diagrams 46
3.6.3 Direct and Indirect Band Gaps 47
3.6.4 Phonons 48
3.7 Summary and Learning Points 49
3.8 Questions 50
3.9 Problems 51

4 Semiconductors as Laser Materials 2: Density of States, Quantum Wells, and Gain 53
4.1 Introduction 53
4.2 Density of Electrons and Holes in a Semiconductor 54
4.2.1 Modifications to Eq. 4.9: Effective Mass 56
4.2.2 Modifications to Eq. 4.9: Including the Band Gap 59
4.3 Quantum Wells as Laser Materials 59
4.3.1 Energy Levels in an Ideal Quantum Well 60
4.3.2 Energy Levels in a Real Quantum Well 62
4.4 Density of States in a Quantum Well 64
4.5 Number of Carriers 66
4.5.1 Quasi-Fermi Levels 67
4.5.2 Number of Holes Versus Number of Electrons 68
4.6 Condition for Lasing 68
4.7 Optical Gain 70
4.8 Semiconductor Optical Gain 71
4.8.1 Joint Density of States 72
4.8.2 Occupancy Factor 73
4.8.3 Proportionality Constant 74
4.8.4 Linewidth Broadening 75
4.9 Summary and Learning Points 77
4.10 Learning Points 77

4.11 Questions . . . 78
4.12 Problems . . . 78

5 Semiconductor Laser Operation . . . 83
5.1 Introduction . . . 83
5.2 A Simple Semiconductor Laser . . . 84
5.3 A Qualitative Laser Model . . . 85
5.4 Absorption Loss . . . 88
5.4.1 Band-to-Band and Free Carrier Absorption . . . 89
5.4.2 Band-to-Impurity Absorption . . . 90
5.5 Rate Equation Models . . . 91
5.5.1 Carrier Lifetime . . . 93
5.5.2 Consequences in Steady State . . . 95
5.5.3 Units of Gain and Photon Lifetime . . . 97
5.5.4 Slope Efficiency . . . 97
5.6 Facet-Coated Devices . . . 99
5.7 A Complete DC Analysis . . . 102
5.8 Summary and Learning Points . . . 105
5.9 Questions . . . 106
5.10 Problems . . . 107

6 Electrical Characteristics of Semiconductor Lasers . . . 111
6.1 Introduction . . . 111
6.2 Basics of *p–n* Junctions . . . 112
6.2.1 Carrier Density as a Function of Fermi Level Position . . . 112
6.2.2 Band Structure and Charges in *p–n* Junction . . . 115
6.2.3 Currents in an Unbiased *p–n* Junction . . . 118
6.2.4 Built-in Voltage . . . 120
6.2.5 Width of Space Charge Region . . . 121
6.3 Semiconductor *p–n* Junctions with Applied Bias . . . 124
6.3.1 Applied Bias and Quasi-Fermi Levels . . . 124
6.3.2 Recombination and Boundary Conditions . . . 126
6.3.3 Minority Carrier Quasi-Neutral Region Diffusion Current . . . 128
6.4 Semiconductor Laser *p–n* Junctions . . . 131
6.4.1 Diode Ideality Factor . . . 131
6.4.2 Clamping of Quasi-Fermi Levels at Threshold . . . 132
6.5 Summary of Diode Characteristics . . . 133
6.6 Metal Contact to Lasers . . . 133
6.6.1 Definition of Energy Levels . . . 133
6.6.2 Band Structures . . . 135
6.7 Realization of Ohmic Contacts for Lasers . . . 140

6.7.1 Current Conduction Through a Metal–Semiconductor Junction: Thermionic Emission . . . 141
6.7.2 Current Conduction Through a Metal–Semiconductor Junction: Tunneling Current . . . 142
6.7.3 Diode Resistance and Measurement of Contact Resistance . . . 143
6.8 Summary and Learning Points . . . 145
6.9 Questions . . . 146
6.10 Problems . . . 147

7 The Optical Cavity . . . 151
7.1 Introduction . . . 151
7.2 Chapter Outline . . . 152
7.3 Overview of a Fabry-Perot Optical Cavity . . . 153
7.4 Longitudinal Optical Modes Supported by a Laser Cavity . . . 154
7.4.1 Optical Modes Supported by an Etalon: The Laser Cavity in 1D . . . 154
7.4.2 Free Spectral Range in a Long Etalon . . . 156
7.4.3 Free Spectral Range in a Fabry-Perot Laser Cavity . . . 158
7.4.4 Optical Output of a Fabry-Perot Laser . . . 160
7.4.5 Longitudinal Modes . . . 161
7.5 Calculation of Gain from Optical Spectrum . . . 162
7.6 Lateral Modes in an Optical Cavity . . . 165
7.6.1 Importance of Lateral Modes in Real Lasers . . . 166
7.6.2 Total Internal Reflection . . . 167
7.6.3 Transverse Electric and Transverse Magnetic Modes . . . 169
7.6.4 Quantitative Analysis of the Waveguide Modes . . . 170
7.7 Two-Dimensional Waveguide Design . . . 175
7.7.1 Confinement in Two Dimensions . . . 175
7.7.2 Effective Index Method . . . 175
7.7.3 Waveguide Design Targets for Lasers . . . 178
7.8 Summary and Learning Points . . . 178
7.9 Questions . . . 180
7.10 Problems . . . 180

8 Laser Modulation . . . 183
8.1 Introduction: Digital and Analog Optical Transmission . . . 183
8.2 Specifications for Digital Transmission . . . 184
8.3 Small Signal Laser Modulation . . . 186

8.3.1 Measurement of Small Signal Modulation 187
8.3.2 Small Signal Modulation of LEDs 188
8.3.3 Rate Equations for Lasers, Revisited 190
8.3.4 Derivation of Small Signal Homogenous Laser Response 193
8.3.5 Small Signal Laser Homogenous Response 195
8.4 Laser AC Current Modulation 196
8.4.1 Outline of the Derivation 196
8.4.2 Laser Modulation Measurement and Equation 197
8.4.3 Analysis of Laser Modulation Response 200
8.4.4 Demonstration of the Effects of τ_c 203
8.5 Limits to Laser Bandwidth 203
8.6 Relative Intensity Noise Measurements 205
8.7 Large Signal Modulation 207
8.7.1 Modeling the Eye Pattern 207
8.7.2 Considerations for Laser Systems 209
8.8 Summary and Conclusions 210
8.9 Learning Points 210
8.10 Questions 212
8.11 Problems 212

9 Distributed Feedback Lasers 215
9.1 A Single-Wavelength Laser 215
9.2 Need for Single-Wavelength Lasers 216
9.2.1 Realization of Single-Wavelength Devices 218
9.2.2 Narrow Gain Medium 218
9.2.3 High Free Spectral Range and Moderate Gain Bandwidth 219
9.2.4 External Bragg Reflectors 220
9.3 Distributed Feedback Lasers: Overview 222
9.3.1 Distributed Feedback Lasers: Physical Structure 222
9.3.2 Bragg Wavelength and Coupling 224
9.3.3 Unity Round Trip Gain 225
9.3.4 Gain Envelope 225
9.3.5 Distributed Feedback Lasers: Design and Fabrication 226
9.3.6 Distributed Feedback Lasers: Zero Net Phase 228
9.4 Experimental Data from Distributed Feedback Lasers 231
9.4.1 Influence of κ on Threshold Current and Slope Efficiency 231
9.4.2 Influence of Phase on Threshold Current 232

9.4.3 Influence of Phase on Cavity Power Distribution and Slope . . . 232
9.4.4 Influence of Phase on Single-Mode Yield . . . 233
9.5 Modeling of Distributed Feedback Lasers . . . 237
9.6 Coupled Mode Theory . . . 240
9.6.1 A Graphical Picture of Diffraction . . . 240
9.6.2 Coupled Mode Theory in Distributed Feedback Laser . . . 241
9.6.3 Measurement of κ . . . 246
9.7 Inherently Single-Mode Lasers . . . 247
9.8 Other Types of Gratings . . . 249
9.9 Learning Points . . . 250
9.10 Questions . . . 251
9.11 Problems . . . 251

10 Assorted Miscellany: Dispersion, Fabrication, and Reliability . . . 255
10.1 Introduction . . . 255
10.2 Dispersion and Single Mode Devices . . . 256
10.3 Temperature Effects on Lasers . . . 258
10.3.1 Temperature Effects on Wavelength . . . 259
10.3.2 Temperature Effects on DC Properties . . . 259
10.4 Laser Fabrication: Wafer Growth, Wafer Fabrication, Chip Fabrication, and Testing . . . 263
10.4.1 Substrate Wafer Fabrication . . . 263
10.4.2 Laser Design . . . 264
10.4.3 Heterostructure Growth . . . 265
10.5 Grating Fabrication . . . 267
10.5.1 Grating Fabrication . . . 267
10.5.2 Grating Overgrowth . . . 269
10.6 Wafer Fabrication . . . 270
10.6.1 Wafer Fabrication: Ridge Waveguide . . . 270
10.6.2 Wafer Fabrication: Buried Heterostructure Versus Ridge Waveguide . . . 272
10.6.3 Wafer Fabrication: Vertical Cavity Surface-Emitting Lasers (VCSELS) . . . 274
10.7 Chip Fabrication . . . 276
10.8 Wafer Testing and Yield . . . 278
10.9 Reliability . . . 279
10.9.1 Individual Device Testing and Failure Modes . . . 279
10.9.2 Definition of Failure . . . 281
10.9.3 Arrhenius Dependence of Aging Rates . . . 282
10.9.4 Analysis of Aging Rates, FITS, and MTBF . . . 282

10.9.5 Electrostatic Discharge and Electrical Overstresses . . . 285
10.9.6 Optical Overstress and Snap Test . . . 286
10.10 Design for … . . . 286
10.10.1 Design Tools . . . 287
10.10.2 Design for High Speed Directly Modulated Lasers . . . 287
10.10.3 Design for High Power . . . 288
10.10.4 Design for Low Linewidth . . . 288
10.10.5 Design Over Temperature . . . 289
10.11 Summary and Learning Points . . . 289
10.12 Questions . . . 290
10.13 Problems . . . 291

11 Laser Communication Systems I: Amplitude Modulated Systems . . . 293
11.1 Introduction . . . 293
11.2 Evolution of Optical Speed . . . 294
11.3 Evolutionary Changes . . . 294
11.4 Multiplexing . . . 296
11.4.1 Wavelength Division Multiplexing . . . 296
11.4.2 Wavelength Division Multiplexing and Demultiplexing . . . 298
11.4.3 Optical Add Drop Multiplexors . . . 299
11.5 Overview of Amplitude-Modulated Communication . . . 299
11.5.1 Definitions for Amplitude Modulation Formats . . . 300
11.5.2 Bits Versus Symbols . . . 300
11.5.3 Pulse Amplitude Modulation . . . 302
11.6 External Modulation . . . 303
11.6.1 Quantum-Confined Stark Effect . . . 304
11.6.2 Absorption Modulation Through the Quantum-Confined Stark Effect . . . 305
11.6.3 Mach–Zehnder Modulator from Electooptic Materials . . . 307
11.6.4 Phase Shifting with Plasma Effect . . . 310
11.7 Laser Linewidth . . . 312
11.7.1 Inherent Laser Linewidth . . . 312
11.7.2 Linewidth Enhancement Factor . . . 313
11.8 Direct Detection Receivers . . . 315
11.9 Summary and Learning Points . . . 317
11.10 Questions . . . 318
11.11 Problems . . . 319

12 Coherent Communication Systems 323
12.1 Introduction 323
12.2 Phasor Representation of Light 324
12.2.1 Reminder: Phasor Representation of Electrical Signals 324
12.2.2 Phasor Representation of Optical Signals 324
12.3 Phasor Descriptions of Coherent Optical Transmission 326
12.3.1 Binary (and More) Phase Shift Keying 326
12.3.2 Differential Phase Shift Keying 326
12.3.3 Quadrature Amplitude Modulation 328
12.3.4 Polarization Division Multiplexing 329
12.3.5 Polarization-Maintaining Fiber 331
12.4 Coherent Optical Transmitters 331
12.4.1 Binary (or More) Phase Shift Keying Transmitter 332
12.4.2 Quadrature Amplitude Modulation 333
12.5 Receivers 334
12.5.1 Reference Signal 334
12.5.2 Balanced Photodiode 334
12.5.3 A Full Coherent System 337
12.6 Coherent Transmission in Context 338
12.6.1 Comparison of Coherent and Incoherent (Amplitude Shift Keying) Systems 338
12.6.2 Communication Formats 339
12.7 Limits to Transmission Distance in Optical Systems 339
12.7.1 Optical Signal-to-Noise Ratio 340
12.7.2 Eye Diagram-Based Signal-to-Noise Ratio 341
12.7.3 Bit Error Rate Versus Transmission Format and Signal-to-Noise Ratio 342
12.8 Noise Sources 344
12.8.1 Relative Intensity Noise 345
12.8.2 Shot Noise 345
12.8.3 Erbium-Doped Fiber Amplifier Noise 346
12.8.4 Thermal Johnson Noise 348
12.8.5 Combination of Noise Sources 349
12.8.6 Other Noise Sources 350
12.9 Final Words 350
12.10 Summary and Learning Points 351
12.11 Questions 353
12.12 Problems 353

Index 355

Introduction: The Basics of Optical Communications

1

Begin at the beginning and go on till you come to the end: then stop.
—Lewis Carroll, *Alice in Wonderland*

Abstract

In this chapter, the motivation for the study of semiconductor lasers (optical communications) is introduced, and the outline of the book described.

1.1 Introduction

It is very difficult to fit a subject like semiconductor laser for optical communications into a single book and have it remain accessible. It spans an enormous range of areas, including optics, photonics, solid-state physics, and electronics, each of which is (by itself) worthy of several textbooks. The objective here is to present semiconductor lasers in a way that is both accessible and interesting to advanced undergraduate students and to first-year graduate students. The target audience for this book is someone who is potentially interested in careers in semiconductor lasers, and the decision of what topic to cover is driven both by the importance of the topic and how fundamental it is to the whole field. We aim to make the reader very comfortable with both the scientific and engineering aspects of this discipline.

Before we leap into the technical details of the subject of semiconductor lasers in communications, it is wise to take a step back to appreciate both the historical and technological significance of these devices in optical communications and the need for semiconductor lasers for light sources in optical communication.

Finally, at the end of the chapter, we would like to introduce the reader to what a semiconductor laser looks like and describe how the book is organized.

© Springer Nature Switzerland AG 2020

D. J. Klotzkin, *Introduction to Semiconductor Lasers for Optical Communications*,
https://doi.org/10.1007/978-3-030-24501-6_1

1.2 Introduction to Optical Communications

1.2.1 The Basics of Optical Communications

Optical communications by itself have a long history. Modern optical communications based on lasers and optical fibers are incredibly attractive communications solution for the fundamental and technological reasons listed in Table 1.1.

The last point is the key advertisement for semiconductor lasers in optical communication. Long ago, Paul Revere used lanterns to signal the arrival and mode of transport of the British invaders. Those lanterns are black-body light sources formed by heat, producing incoherent light in a spectrum of wavelengths and propagating through a turbulent, lossy atmosphere. Nonetheless, information was conveyed for miles. To truly take advantage of the amazing properties of light, and transmit light for hundreds of miles, a convenient, single-wavelength coherent source is needed, along with a very clear, lossless waveguide. The answer to the first requirement is a semiconductor laser.

The basis of fiber optic communications is pulses of light created by lasers transmitted for many hundreds or thousands of miles over optical fiber. An enormous amount of information can be transmitted over each fiber. Light of different wavelengths can transmit without affecting each other, and light at each wavelength can transmit data up to many gigabits/second.

The vast majority of these bits are generated by semiconductor lasers, which are one of the most useful inventions of the second half of the twentieth century. The first coherent emission from semiconductors was demonstrated in 1958 by a group led by Robert Hall. The first modern double heterostructure laser was proposed by Herbert Kroemer and ended up earning him and Zhores I. Alferov the 2000 Nobel Prize for 'developing semiconductor heterostructures used in high-speed- and opto-electronics' (http://www.nobelprize.org)[1]. Jack S. Kilby also received the 2000 Nobel Prize for 'his part in the invention of the integrated circuit.'

Fiber optic technology enables billions and billions of bits to flow seamlessly and uninterrupted from one side of the world to the other.

The building blocks for this optical communication network are shown in Fig. 1.1. Figure 1.1a shows coils of optical fiber, demonstrating the portability and compactness of this flexible and convenient routable waveguide. Figure 1.1b–d shows several types of optical communication packages, from Fig. 1.1b, a single semiconductor laser transmitter already fiber coupled, to Fig. 1.1c, an integrated transceiver which connects a digital electrical side to an optical side, with the drive circuitry as part of the package, to Fig. 1.1d, a very low-cost TO-can laser package. In all of these, the electrical signal is modulated onto the light, which is connected to an optical fiber. Miles of this are routed under the ground, and enormous bandwidth is available everywhere.

[1] An interesting story: according to Herbert Kroemer, he first wrote up this idea and submitted it as a paper to the journal *Applied Physics Letters,* and it was rejected. Sometimes important ideas are difficult to recognize!

Table 1.1 Advantages of optical communications

Light has enormous bandwidth	As an electromagnetic wave with a frequency in the hundreds of THz, a lot more information can be carried with light than can be carried on electromagnetic waves of lower frequency in conventional electromagnetic spectrum
Light is easily guided	Flexible and very low-loss waveguides (glass fibers) have been invented that allow these pulses of light to be routed just like electrical signals
Light can be easily detected and generated	The best wavelengths for transmission can be easily generated and detected with semiconductor devices, and these sources and detectors can be economically fabricated

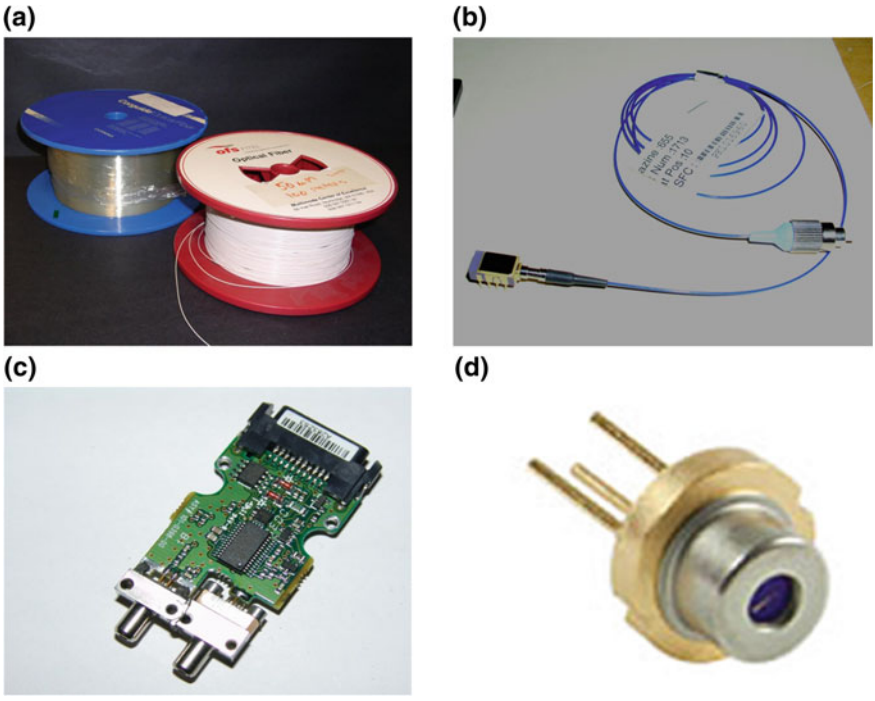


Fig. 1.1 **a** An unjacketed coil of optical fiber containing 20 km (12 miles) of fiber and a jacketed coil of fiber containing 100 m; **b** a semiconductor laser transmitter showing electrical inputs with an optical output; **c** an integrated gigabit laser transceiver module, with a digital electrical interface and optical output and input side, from Wikipedia (https://en.wikipedia.org/wiki/Fiber-optic_communication, current 1/2019); **d** a simple and economical TO-can laser package, from Thorlabs, picture used by permission

The growth of this use of bandwidth can be seen in Fig. 1.2. As of 2006, the amount of digital data was doubling about every ~2 years; now (2019) it has slowed to a point where it is doubling only every three years. This is a prodigous growth rate!

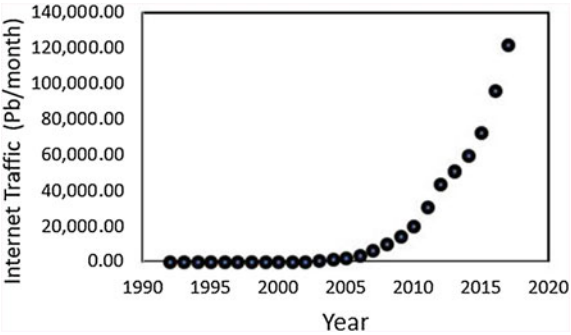


Fig. 1.2 Worldwide growth of internet bandwidth traffic. Data from https://en.wikipedia.org/wiki/Internet_traffic, current 12/2019

To give a sense of the power of fiber optic transmission, the demonstrated bandwidth that can be transmitted over a single optical fiber is about 100 Tb/s. There is tremendous bandwidth capacity in optical fiber, and most optical fibers are drastically underutilized.

1.2.2 A Remarkable Coincidence

Optical communications are based on the transmission of light pulses through optical fiber. It owes its remarkable utility to a very fortunate coincidence and a fortuitous invention. The coincidence is illustrated in Fig. 1.3. The invention was made by Maurer, Schultz, and Keck at Corning when they first demonstrated 'low' (20 dB/km) loss fiber at Corning in 1970.

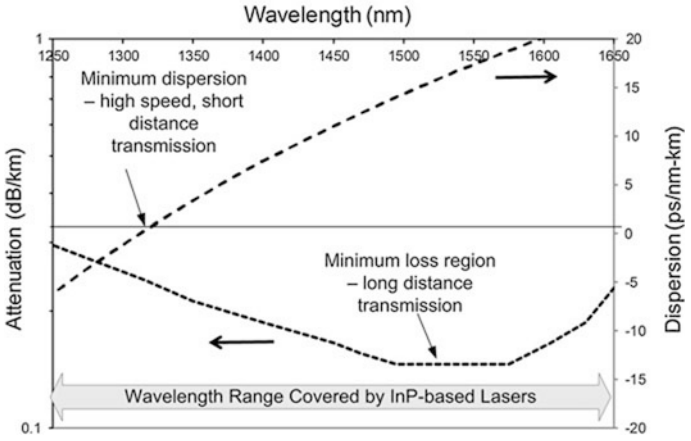


Fig. 1.3 Fiber attenuation and dispersion versus wavelength, over the bandwidth range covered by InP-based semiconductor lasers most often used for telecommunications lasers

Figure 1.3 shows the optical loss in current state-of-the-art single-mode glass fiber, in units of dB/km. Modern Corning SMF-28 optical fiber has a loss minimum about 0.2 dB/km at a wavelength around 1550 nm. If the objective is to transmit power as far as possible, this lowest loss wavelength of 1550 nm is the best choice of wavelength. (For reasons we will talk about later, the low-dispersion window around 1310 nm is also highly desirable).

Where do the light sources to transmit this information come from? Semiconductor lasers are made with semiconductors, and semiconductors have a natural property, called the band gap, which controls the wavelength of light they can emit. Figure 1.3 also indicates the broad range of wavelengths that can be generated or detected by InP-based semiconductors used as both sources and detectors. It happens that wavelengths around 1300 and 1550 nm are easily accessible by making heterostructures of the different semiconductors appropriately.

Hence, sources that create light in the low-loss region of glass (at a wavelength around 1550 nm) can be easily fabricated in semiconductors. Semiconductor lasers and light-emitting diodes are marvelously convenient sources of light—they are small, simple to make and inexpensive and can take advantage of all the expertise and background that has grown up around fabricating semiconductors for standard electronics. This fortunate match between conveniently fabricated light source and the particular wavelength needed has led to the tremendous growth and importance of this technology. Without these convenient light sources, and availability of an excellent waveguide, other technology may have been chosen as the technology of choice for communications.

An excellent overview of extraordinarily rapid growth of fiber optic technology is given in the book *City of Light: The Story of Fiber Optics*, by Jeff Hecht.

1.2.3 Optical Amplifiers

The third leg of this technology for optical communication is the invention of the erbium-doped fiber amplifier (EDFA) in 1986 or 1987. Even though the loss in optical fiber had been reduced to a point where 100 km transmission does not require amplification, amplification is required for distances greater than 100 km. For global connectivity, a convenient way to *optically* amplify these signals was needed. The alternative of receiving the optical signal, translating it back to electrical data and then retransmitting optically every 100 km was a serious drawback to the widespread adoption of optical communications.

The EDFA is a device that can directly amplify all the light signals in a fiber, at any practical speed, without converting them back into electrical signals and regenerating them. With EDFAs, the limitation to long-distance transmission becomes dispersion or overall signal-to-noise ratio. Depending on the transmitter, that distance could be 600 km or even longer.

1.2.4 A Complete Technology

This collection of interlocking technologies (along with others that we have not mentioned, such as dispersion-compensated fiber and optical switching techniques) has enabled this entire field to take off and blossom. Low-loss waveguides and optical amplifiers enabled precise routing of transmission of these signals over tremendous distances—since semiconductors are convenient sources and also receivers of the light signal they take advantage of the vast semiconductor manufacturing infrastructure. Voltaire would say (truly) that we are optically in 'the best of all possible worlds.'

1.3 A Picture of Semiconductor Lasers

Before we introduce the mechanics and physics of semiconductor lasers, it is useful to convey an overall broad picture of what they are. The details in this overview here will be covered in subsequent chapters.

Semiconductor lasers start out as pieces of semiconductor wafer (let us say an InP base) with various other layers deposited on it. This epitaxial base wafer is (as close as engineering can get it) a perfect crystal. Seen in visible light, a polished wafer is an excellent mirror. At wavelengths below the band gap, in the far infrared, the wafer appears as transparent as a piece of extra-clean window glass in ordinary visible light.

The wafer is processed by depositing more layers on it and finally mechanically breaking or 'cleaving' it into thin strips of laser bars. Each of these laser bars has many tens of lasers on it. These lasers are then broken into individual laser devices, each typically about 0.5 mm long (about the same as a large grain of rice), and mounted and packaged. Testing and packaging these devices is typically much harder than testing or packaging electrical devices, since the cleaving (breaking apart) of the wafer is what forms the surface of the cavity mirror, and that must be kept to perfect optical smoothness. The final packaged device will be coupled to an optical fiber, which also takes precision mechanical handling (compare that to a microprocessor, which only needs electrical contact to each of the electrical pads)!

These aspects of laser semiconductors will all be covered in detail in subsequent chapter. It is useful though to see something before discussing the physics behind it, and so, we partly interrupt the flow of narrative to now show a semiconductor laser.

Figure 1.4 shows some of the stages of development of a semiconductor laser, from a wafer, to a bar, to a chip, to a sub-mount. That sub-mount will be eventually packaged as shown in Fig. 1.1.

Figure 1.5 shows a close-up view of a typical semiconductor laser. The figure shows the waveguide (here a ridge waveguide device), the semiconductor active region medium (quantum wells), the top and bottom metal contacts (by which current is injected), and the optical mode (the shape of the spot of light in the semiconductor). The secondary electron microscope picture on the right shows the

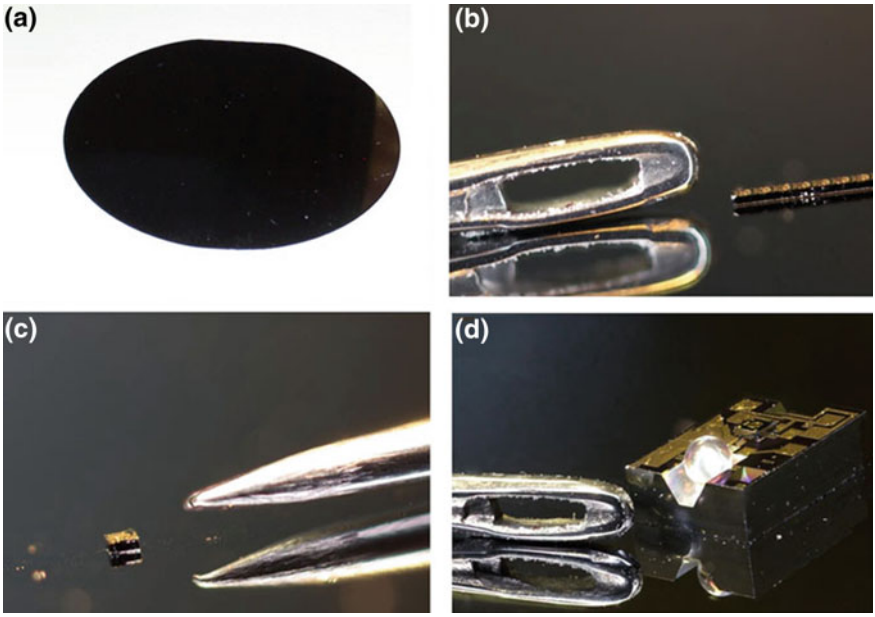


Fig. 1.4 Stages of development in a semiconductor laser. **a** It first starts as an epitaxial wafer, upon which different layers of material are grown, metals are deposited, and various processing steps are made. **b** It is then fabricated through etching, metal deposition, and other microfabrication steps and then separated into individual laser bars as shown in (**b**). **c** Each bar is separated into individual chips, and **d** chips are packed by being soldered to sub-mounts and then coupled into an optical fiber. The scale factor in figures (**b**) and (**c**) is the point of a needle; in (**d**) it is the eye of a needle. The mechanical handling of such small devices is a major part of fabrication of optical transmitters. Each individual laser is packaged separately; potentially ten thousand lasers can be obtained from a single wafer. Photograph credit J. Pitarresi

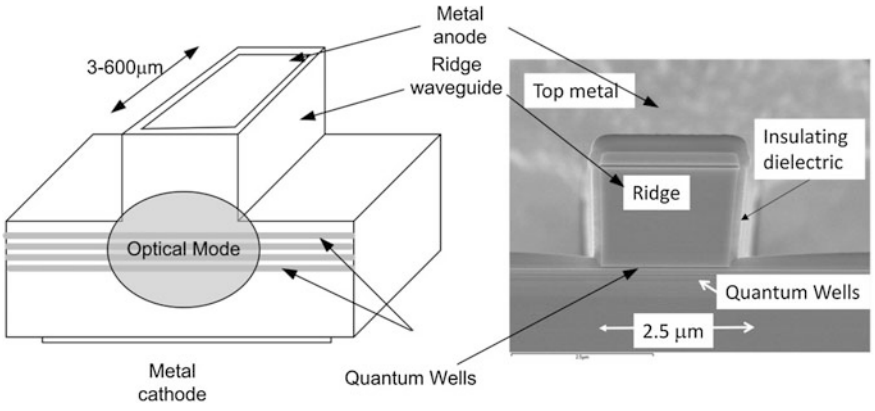


Fig. 1.5 Schematic of a ridge waveguide semiconductor laser, and a picture of the front facet of a ridge waveguide device

actual dimensions of a complete laser—the ridge is typically a few microns wide and tall, and the quantum well area (the 'active region') is about 300 nm or so thick. (Quantum wells, largely the subject of Chap. 4, are thin slabs of material sandwiched by other materials which give the device beneficial properties). The ridge length is around 3–600 μm (about 0.5 mm). (This is only one of several common laser structures. This is called a ridge waveguide—other types will be discussed later in the book.)

In Fig. 1.5, current is injected through the top and bottom, and light is coming out front and back (along the line of the ridge).

1.4 Organization of the Book

In general, topics in this book will be covered in order from most general to most specific. In this first chapter, the motivation for the study of semiconductor lasers and a general introduction to the field of optical communication was presented. Table 1.2 shows the organization of the book by chapter. Chapter 2 will discuss general properties of all lasers made of any material. Chapter 3 will discuss the basics of semiconductors as a lasing medium, including details of the band structure, strained-layer growth, and direct and indirect semiconductors, heterostructures, strain and grown ideal semiconductors, including the band gap, density of states, quasi-Fermi level, and optical gain. Chapter 4 introduces quantitative models of the density of states for both bulk and quantum well systems and discusses the conditions for population inversion.

Chapter 5 ties together the qualitative laser models with measureable performance characteristics, such as slope and threshold current, and describes some of the common experimental metrics used to evaluate laser material. Chapter 6 takes a break from talking about optical and material characteristics and instead talks about the specific electrical characteristics of semiconductor junction lasers, including metal contacts. Chapter 7 discusses the laser as an optical cavity, including design of single-mode waveguide and mode separation in Fabry-Perot cavities.

Chapters 8 and 9 talk more specifically about laser communications, partly issues relevant to directly modulated lasers. Chapter 8 discusses laser modulation and the inherent limitations to semiconductor laser speed. The focus of Chap. 9 is single-wavelength distributed feedback lasers and the inherent variability introduced with a grating and the usual high-reflection/anti-reflection coatings.

Chapter 10 covers a number of other more applied topics, such as dispersion in laser transmission, laser reliability, temperature dependence of laser characteristics, and laser fabrication. Though laser applications and requirements are unique, it also includes a brief section on laser design (new to the second edition).

Chapter 11 (new to the second edition) introduces amplitude modulated optical communication systems. On–off modulation is extended into pulse amplitude modulation, and the physics of methods of external modulation and optical receivers are briefly discussed.

Table 1.2 Organization of the book

Chapters	Topics
1	Introduction to optical communication and to organization of the book
2	Structure and requirements for all lasers, semiconductor, or other materials
3	The ideal semiconductor and quantum wells, heterostructures and strained-layer growth, direct, and indirect band gap
4	Density of states in semiconductor lasing medium, conditions for population inversion, quasi-Fermi levels
5	Connection between laser model and measured characteristics of threshold current and slope efficiency
6	Electrical characteristics of semiconductor lasers. I-V curve, metal connections
7	Optical cavities in semiconductors, and the relationship between gain and cavity. Design of single-mode cavity
8	High-speed properties of semiconductor lasers—rate equation models
9	Single-wavelength lasers; distributed feedback lasers
10	Other miscellaneous topics including fabrication, communication, yield, and reliability; and laser design
11	Directly modulated laser communication systems, symbol and bit rates, pulse amplitude modulation
12	Coherent laser communication systems

Chapter 12 (also new to the second edition) discusses coherent optical communication systems, and the signal-to-noise limits of optical communication formats.

1.5 Summary and Learning Points

A. Optical communications continue to grow rapidly very rapidly driven by enormous growth rate of worldwide bandwidth usage.
B. Basic optoelectronic communication systems consist of semiconductor laser light sources coupled to flexible fiber optic waveguides.
C. These fiber optic waveguides can carry enormous amounts of bandwidth.
D. Optical transmitters and receivers based on semiconductors can be fabricated to transmit at the low-loss and low-dispersion points of optical fiber.
E. Erbium-doped fiber amplifiers can amplify optical signals in fiber without need of regeneration.
F. All of these make optical communications based on lasers a near-perfect communication solution.

1.6 Questions and Problems

Q1.1. What are optical communications?
Q1.2. Why do we use lasers and optical fibers in optical communications?
Q1.3. What are the particular advantages of semiconductor lasers in optical communications?
Q1.4. Identify a few semiconductors on the Periodic Table.
Q1.5. What is an EDFA?
Q1.6. What are typical dimensions of the active region of a semiconductor laser?

The Basics of Lasers

2

But soft, what light through yonder window breaks…
—Shakespeare, Romeo and Juliet

Abstract

In this chapter, the important common elements of all lasers are introduced. Some examples of lasing systems are given to define how these elements are implemented in practice.

2.1 Introduction

Semiconductor lasers are the enabling light source of choice for optical communications. However, the basic principles of operation of semiconductor lasers are shared by all lasers. In this chapter, the requirements for lasing systems and the characteristics of all lasers will be discussed. Specific examples from outside of semiconductor lasers will be used to demonstrate these characteristics, before we focus on the specific mechanics and structure of semiconductor lasers.

2.2 Introduction to Lasers

With an appreciation of the significance and underlying technology of optical communication, we can start to understand the basic process of lasing. In this section, we introduce the fundamental underpinnings of lasing, stimulated emission. Stimulated emission is the idea that under certain conditions a photon can create additional photons of the same wavelength and phase. Lasers are based on this principle and create 'floods' of photons of the same wavelength and phase that constitute laser light.

© Springer Nature Switzerland AG 2020

D. J. Klotzkin, *Introduction to Semiconductor Lasers for Optical Communications*,
https://doi.org/10.1007/978-3-030-24501-6_2

To start to understand stimulated emission, we begin with a description of one of the classical problems of physics, black body radiation.

2.2.1 Black Body Radiation

Black body radiation is the spectrum emitted from a 'black body' (an object without any particular color) as it is heated up. 'Red hot' iron and 'yellow hot' iron are red and yellow because, at the temperature to which they are heated, their emission peak is ~600 or ~550 nm, and they look 'red' or 'yellow.' The surface of the sun is another example of a classical black body. Measurements showed that black bodies emit light at a peak spectral wavelength depending on their temperature, with the amount of emission above and below that wavelength falling off to zero at shorter and longer wavelengths. The peak emission shifted to shorter wavelengths as the temperature of the black body increased. All black bodies at the same temperature emit light of the same spectrum, independent of the material.

In the beginning of the twentieth century, the physics behind the spectrum was a great mystery to early twentieth-century physicists. The shape of the curve was well described by a simple equation first derived by Max Planck,

$$E(v)\mathrm{d}v = \frac{8\pi h v^3}{c^3}\frac{1}{\exp(hv/kT)-1}\mathrm{d}v \tag{2.1}$$

where $E(v)$ is the amount of energy density, in J/m^3/Hz, in each frequency.[1] The theory behind this equation was not understood until quantum mechanics was introduced.

Aside: It is remarkable how powerful and universal this black body spectrum is. Radiation from outer space is difficult to measure on Earth, because the atmosphere absorbs very long wavelengths. The Cosmic Background Explorer (COBE) satellite was sent up to measure the far-infrared black body spectrum above the atmosphere. Shown in Fig. P2.1 is one of the spectra it recorded. The shape fits perfectly to the shape of the spectrum of Eq. 2.1, and from this data, the temperature of the universe could be extracted. It turns out that the universe as a whole is a balmy 2.75 K. This measurement is currently being interpreted as support for the Big Bang theory of the creation of the universe. It was clear this measurable phenomenon was driven by basic physics. The initial theory and discovery of this cosmic background

[1] M. Planck, 'On The Theory of the Law of Energy Distribution in the Continuous Spectrum,' Verhandl. Dtsch. Phys. Ges., 2, 237.

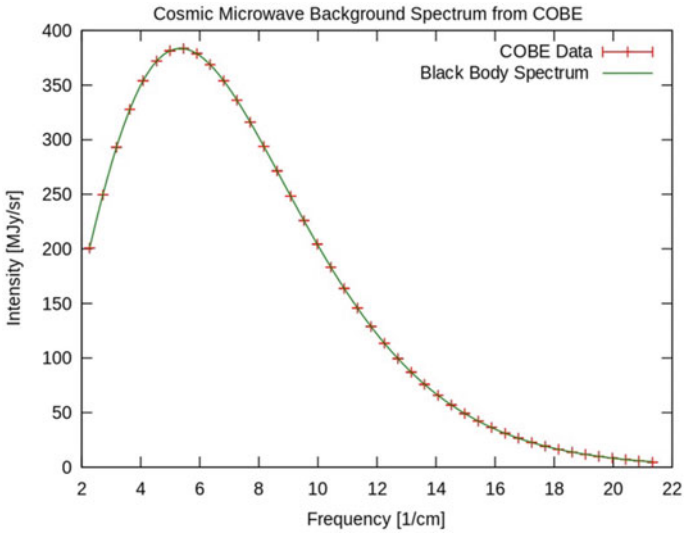


Fig. P2.1 One of the first measurements of the COBE background microwave satellite, showing the use of the optical spectrum of the black body to measure temperature. Image from http://en.wikipedia.org/wiki/File:Cmbr.svg, current 1/2013

```
radiation resulted in Nobel Prizes for Penzias and Wil-
son in 1978; the subsequent measurements by the COBE
satellite resulted in Nobel Prizes for Smoot and Mather.
```

This black body formula can be understood in fundamentally two different ways: (i) a macroscopic, statistical thermodynamics viewpoint, attributed to Planck, and (ii) a microscope rate equation viewpoint, attributed to Einstein. Both views are correct, and both have parallels with semiconductor lasers. The statistical view, involving density of states, is repeated when calculating gain in a semiconductor laser. The rate equation view comes up again when talking about modeling laser DC and dynamic performance. Let us talk about both views in detail.

2.2.2 Statistical Thermodynamics Viewpoint of Black Body Radiation

The viewpoint of statistical thermodynamics, which is fundamentally Planck's view, is that an existing 'state' has a certain probability to be occupied, based on its temperature. As the temperature increases, it becomes more likely that higher energy states are occupied. At a temperature of absolute zero, only the very lowest energy states are occupied: At higher temperatures, the higher-level energy states start to be occupied.

As such, the spectrum is determined by two things: first, the probability distribution function, which determines the likelihood that a state will be occupied based on temperature, and second, the density of states, which is the number of

states that exists at a particular energy in a black body. We will talk about both of these terms in the next sections.

2.2.3 Some Probability Distribution Functions

Let us briefly review probability distribution functions for photons and electrons. A distribution function gives the probability that an existing state will be occupied based on the energy of the state and the temperature of the system. These functions are thermodynamic functions that are applicable to systems in thermal equilibrium at a fixed temperature. Table 2.1 shows a list of the statistical distribution functions and the systems (or particles) to which they apply.

In these functions, E refers to the energy of the state, E_f is a characteristic energy of the system (the Fermi energy) usually used with Fermi–Dirac statistics, and kT is the Boltzmann constant times the temperature (in Kelvin). The constant A in the Bose–Einstein and Maxwell–Boltzmann functions depends on the type of particles but is 1 for photons.

Example: If the Fermi energy of a semiconductor is 1 eV above the valence band, at room temperature, what is the probability that an electronic state 2 eV above the valence band will be occupied?

Solution: The Fermi–Dirac function applies here, but in fact, $E - E_f$ is high enough that all three functions will give the same answer: $\exp(-1\,\mathrm{eV}/0.026\,\mathrm{eV}) = \exp(-40) = 10^{-18}$.

The *Bose–Einstein* distribution function is appropriate for photons, phonons, and particles with integral spin (like protons) and reflects the fact that these particles can have any number of particles in a given state.

The *Fermi–Dirac* function applies to particles which follow the *Pauli exclusion principle* that at most one particle can occupy a given energy state. Let us take this very earliest opportunity to note that this exclusion principle excludes more than one particle from each quantum *state*, not from each energy level. A quantum state is a set of quantum numbers that describes a particle. Many situations have multiple states with the same energy that have different sets of quantum numbers, such as the sub-levels of p-orbital of an atom. These states are called *degenerate* in energy.

Table 2.1 Distribution functions $P(E)\mathrm{d}E$

Distribution function name	Function	Applies to
Bose–Einstein	$\frac{1}{A\exp^{(E/kT)}-1}$	Bosons: photons and protons and spin-1 particles
Fermi–Dirac	$\frac{1}{\exp^{((E-E_f)/kT)}+1}$	Electrons and other spin ½ particles
Maxwell–Boltzmann	$A\exp^{-(E/kT)}$	All particles at high temperatures

This distribution function is only part of the story. The population of electrons present at any given energy depends on the number of states at that energy. The bandgaps of semiconductors are devoid of states, because of their particular crystalline arrangement. In order to determine the population of photons, we have to derive the density of states, or the number of photon states is available to be occupied at any given energy.

2.2.4 Density of States

In order to apply the distribution functions, a state must exist. These states are allowed solutions of the Schrodinger equation for a particular physical situation or potential.

The calculation of the density of states in black body is best illustrated by an example. Let us proceed to consider the density of photon states for a cubic black body with length L per side and calculate what the density of states per unit energy $D(E)\mathrm{d}E$ is. A picture of a cubic black body volume is shown below. The 'volume' is considered to be macroscopic and much larger than the wavelength of the photons corresponding to this energy.

An intuitive picture suggests that for a given volume, there should be many more short-wavelength, high-energy photons, per volume than long-wavelength, low-energy photons.

The conventional approach here is to pick an electromagnetic boundary condition that confines photons within the black body, and allow only wavelengths that are integral fractions of the cubic length L. For example, wavelengths of $\lambda_x = L$ are allowed, and wavelengths of $\lambda_x = L/2$ are allowed, but a wavelength of $\lambda_x = 0.8L$ is not allowed. The same applies to wavelengths in the other two directions, λ_y and λ_z as shown in Fig. 2.2.

Let us calculate the number of these allowed photon states that exist as a function of energy in a black body.

It is easier to analyze this problem in what is called *reciprocal space*, in which the propagation constants k rather than the wavelengths are considered. If the wavelength is λ_x, the propagation constant $k_x = 2\pi/\lambda_x$. This relationship is true for wavelengths of the components of the photon in each of the three directions, as well as the scalar wavelength of the photon and the amplitude of k.

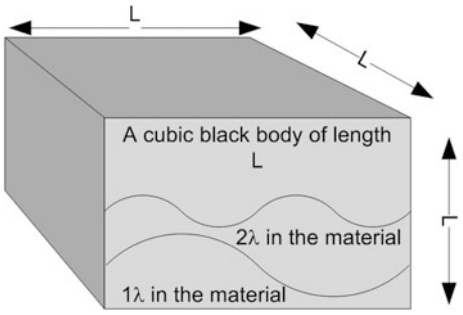


Fig. 2.2 A cubic black body of macroscopic size

We are going to write the relationship between λ and k in two ways (shown below): the first between the vector x, y, z components of k, and the second between the magnitude of k and magnitude of λ. The magnitudes of k and λ are related to their magnitudes in the three orthogonal directions as shown.

$$\begin{aligned} k_{x,y,z} &= \frac{2\pi}{\lambda_{x,y,z}} \\ k &= \sqrt{k_x^2 + k_y^2 + k_z^2} \\ \frac{1}{\lambda} &= \sqrt{\frac{1}{\lambda_x^2} + \frac{1}{\lambda_y^2} + \frac{1}{\lambda_z^2}} \\ \lambda &= \frac{2\pi}{k} \end{aligned} \tag{2.2}$$

The simplest way to understand the propagation constants is to consider them as reciprocals of the wavelength λ. The product of wavelength and propagation constants is a full cycle, 2π. If the wavelength halves, the propagation constant doubles. Writing the allowed wavelengths and propagation constants in terms of the boundary conditions above gives a picture of the spacing of the allowed propagation constants.

The allowed wavelengths are integral fractions of the cavity length, and so, the allowed propagation constants are integral multiples of the fundamental propagation constant, $2\pi/\lambda$, as shown in the expressions below.

$$\begin{aligned} \lambda_{\text{allowed}-x,y.z} &= \frac{L}{m_{x,y.z}} \\ k_{\text{allowed}-x,y,z} &= \frac{m_{x,y,z} 2\pi}{L} \end{aligned} \tag{2.3}$$

These allowed propagation constants form a set of evenly spaced grid points in the reciprocal space plane, as shown below in 2D (x and y). Any point represents a valid propagation constant of a photon, and k-values between the points cannot exist in a black body.

The vector $\mathbf{k}$, having k_x, k_y, and k_z components, gives the propagation direction, and the quantization condition (Eq. 2.3) is independently fulfilled in each direction.

Figure 2.3 shows the picture of allowed k-states in x and y. Using this diagram, and the probability distribution function for photons, we calculate the density of photons at a given frequency (the black body spectrum, Eq. 2.1). What is the number of states at a given energy as a function of the optical frequency v ($N(v)\mathrm{d}v$)?

First, we realize that by Plank's formula, $E = hv$, the optical frequency or wavelength λ equivalently specifies the energy.

$$E = hv = \frac{hc}{\lambda} = \frac{hck}{2\pi} = \hbar ck \tag{2.4}$$

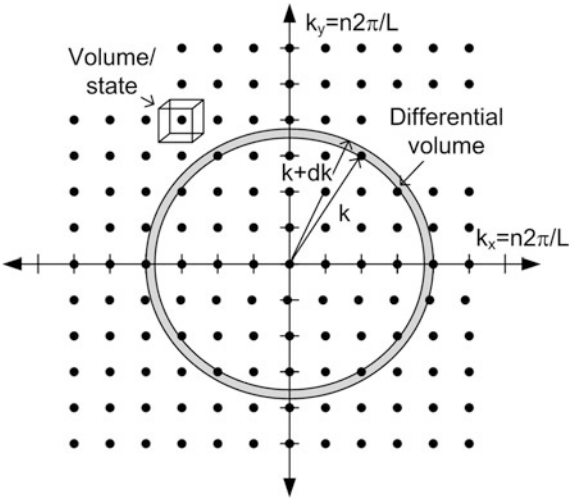


Fig. 2.3 A picture of the allowed points in *k*-space, illustrating the calculation of the 'density of states' of the photon modes in a black body. The picture shows the *x*–*y* plane in *k*-space, but allowed points are also equally spaced in *z*

Even though **k** is a vector as above, the k in this expression is the scalar magnitude of $\boldsymbol{k}$. In the picture above, anything with the same magnitude (shown by the circle) has the same energy. *Calculating the density of states is equivalent to calculating the density of points of a circle of radius k.*

The picture above, for clarity, is actually a 2D picture slice of the 3D system. We are going to carry through the derivation in 3D in which there are three dimensions of allowed propagation vectors, in x, y, and z. The procedure we follow is to calculate the differential volume in a thin slab of fixed radius dk and then divide by the volume per point to get the number of points in that volume. We find that the differential volume for a 3D segment is

$$V(k)\mathrm{d}k = 4\pi k^2 \mathrm{d}k. \tag{2.5}$$

The density of points as a function of k, $D_p(k)$, is given by this volume divided by the density of states in k-space, which is 1 state per $(2\pi/L)^3$ volume, or

$$D_p(k)\mathrm{d}k = \frac{4\pi k^2 \mathrm{d}k}{\left(\frac{2\pi}{L}\right)^3} = \frac{L^3 k^2}{2\pi^2}\mathrm{d}k \tag{2.6}$$

Finally, the relationship between energy and k is best expressed as follows (and substituted into the above)

$$\begin{aligned} E &= \hbar c k \quad \mathrm{d}E = \hbar c\, \mathrm{d}k \\ k &= E/\hbar c \quad \mathrm{d}k = \mathrm{d}E/\hbar c. \end{aligned} \tag{2.7}$$

Substituting into the above expression, we obtain

$$D_p(E)\mathrm{d}E = \frac{4\pi E^2 \mathrm{d}E}{\hbar^3 c^3 \left(\frac{2\pi}{L}\right)^3} = \frac{L^3 E^2 \mathrm{d}E}{2\pi^2 \hbar^3 c^3}. \tag{2.8}$$

Considering the density of states per fixed real space volume, L^3, gives us the nearly final result for the density of points in k-space (D_p) equal to

$$D_p(E)\mathrm{d}E = \frac{E^2 \mathrm{d}E}{2\pi^2 \hbar^3 c^3}\ \mathrm{cm}^{-3} \tag{2.9}$$

A final factor of two has to be multiplied to the expression above to give the density of photon states. Each state, in addition to direction, has a polarization. The polarization can be uniquely specified with two orthogonal polarization states, and as a result the density of state is doubled and the final expression for total density of states, $D(E)$, is

$$D(E)\mathrm{d}E = \frac{E^2 \mathrm{d}E}{\pi^2 \hbar^3 c^3}\ \mathrm{cm}^{-3} \tag{2.10}$$

We have derived this equation in such detail because this will echo the discussion of density of states in an atomic solid, and the very same principles will be used to write down a 'density of states' for electrons and holes in exotic quantum confined structures, like quantum wells (a 2D slab), quantum wires (a 1D line), or quantum dots (small chunks of material with dimensions comparable to atomic wavelength).

Let us make some comments about this derivation, so far. First, there is a key role about the *dimensionality* of the solid. The expression for 'differential volume' contains k^2, which is what leads to the quadratic dependence of $D(E)$ on E. When we start discussing atomic solids, particularly 2D quantum wells (QWs), 1D quantum wires, and 0D quantum dots (QDs), this dimensionality will be different and the density of states will have a different dependence on energy.

Second, let us emphasize again what the term 'density of states' means. It means only the number of states with the same energy, but not with the same quantum numbers. In a black body, for example, there are red photons radiating in all directions, with different quantum numbers $k_{x,y,z}$ but the same wavelength (energy). Density of states measures the number of photons with that red energy or wavelength.

Third, looking back, there is a key assumption about the electromagnetic boundary condition perfectly confining the photons, which is only reasonable and not rigorous.

2.2.5 Spectrum of a Black Body

Having discussed density of states and calculated the density of states in a black body, we now talk about the spectrum of a black body. The statistical thermodynamics way of looking at it is simple: Multiply the density of states by the distribution function (giving the probability that the existing state is occupied) to determine the occupation or emission spectrum. In this case, written as a function of energy, the number of photons $N(E)$ at that energy is:

$$N(E)\mathrm{d}E = \frac{1}{\exp(E/kT) - 1}\frac{E^2\mathrm{d}E}{\pi^2\hbar^3c^3}\ \mathrm{cm}^{-3} \tag{2.11}$$

Or as a function of energy $\rho(E)$ (energy/cm^3), it simply gets multiplied by another E to obtain

$$\rho(E)\mathrm{d}E = \frac{1}{\exp(E/kT) - 1}\frac{E^3\mathrm{d}E}{\pi^2\hbar^3c^3}\ \mathrm{cm}^{-3} \tag{2.12}$$

It is left as an exercise to the student to substitute back in $E = h\nu$ and obtain Planck's black body spectra, Eq. 2.1!

All of this discussion should be relatively familiar. We now want to look at this problem in a slightly different way and see what insights we can get in particular about lasing.

2.3 Black Body Radiation: Einstein's View

The preceding discussion about black bodies introduced (or reminded) the reader of distribution functions and density of states, and both of these concepts will reappear again in the context of semiconductor lasers. However, let us consider a microscope rate equation view, attributed to Einstein, which considers the processes that the photons undergo to maintain that distribution.

Let us consider for a moment the 'sea' of electrons and atoms in a metal which constitute a black body. At any given moment, some number of photons are being absorbed by the metal with the electrons rising to a higher energy level, and some other photons are being emitted as the electrons relax to a lower energy level. For a black body (which is a temperature-controlled, thermodynamic system) at a fixed temperature, these rates of up and down transitions have to be the same for the black body to be in equilibrium. The rate of photons being absorbed has to equal the rate of photons being emitted.

What Einstein postulated was three separate processes which go on in a black body:

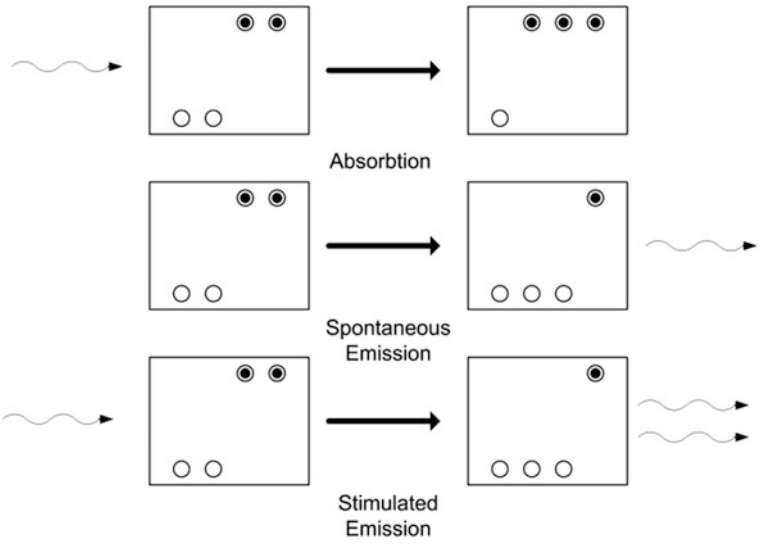


Fig. 2.4 Three processes which occur in a black body and are in equilibrium. Top, absorption; middle, spontaneous emission; and bottom, stimulated emission. The dark circles represent excited states at energy E_2, while the open circles represent unexcited (ground) states at lower energy E_1

(1) Absorption, in which a photon is absorbed by the material and the material (or electron in the material) is left in an excited state;
(2) Spontaneous emission, in which the material or photon relaxes to a lower energy state and a photon is emitted, without the influence of another photon;
(3) Stimulated emission, in which the material or electron relaxes to another energy state and a photon is emitted, when *stimulated by another photon.*

These three processes are illustrated in Fig. 2.4.

It is this last process which is the process responsible for lasing and which we will discuss in much detail. It is likely to be unfamiliar to the student. The proof that in fact it is a valid physics process, as valid as gravitation, will be found in the equivalence of this model with the statistical thermodynamic model of black body emission, when this mechanism is considered.

Let us now proceed to establish the correspondence between these two models. In equilibrium, the rates of the excitation and relaxation processes must be equal. Let us go ahead and postulate the following linear model for the relative rates.

The processes pictured in Fig. 2.5 can be written down conceptually, in equilibrium, as

$$AN_2 + B_{21}N_2N_p(E) = B_{12}N_1N_p(E) \tag{2.13}$$

where N_2 and N_1 are the fraction of the populations in the states N_2 with energy E_2 and N_1 with energy E_1, respectively, $N_p(E)$ is the photon density as a function of

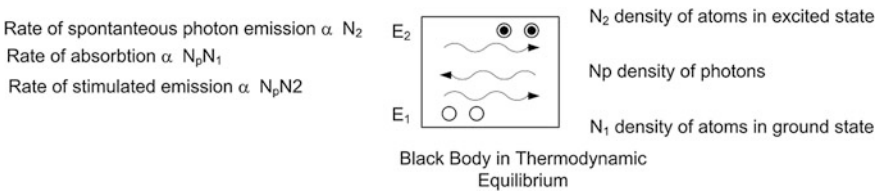

Rate of spontanteous photon emission α N2
Rate of absorbtion α NpN1
Rate of stimulated emission α NpN2
E2
E1
N2 density of atoms in excited state
Np density of photons
N1 density of atoms in ground state
Black Body in Thermodynamic Equilibrium

Fig. 2.5 Processes which go on in a black body, pictured as a collection of photons and excited/unexcited electronic states

energy $E = E_2 - E_1$, A is a linear proportionality coefficient for the rate of absorption, and B_{12} and B_{21} are the linear coefficients for the rates of stimulated emission and absorption, respectively. We include one more physical fact that the populations in state N_1 and state N_2 are in thermodynamic equilibrium, as

$$\frac{N_2}{N_1} = \exp(-(E_2 - E_1)/kT) \tag{2.14}$$

with E_2 and E_1 the energy of the states. With these facts, it is possible to show that the black body spectrum, $N_p(E)$, is the same as that derived earlier *if the two Einstein B* coefficients for stimulated emission and absorption are equal (and we will henceforth write them just as B). This will be left as an exercise for the student (see Problem P2.2)!

2.4 Implications for Lasing

The sense of lasing is of a monochromatic and in-phase beam of light. The process of stimulated emission is one in which a single photon stimulates the emission of another photon, which stimulates additional photons (still in phase at the same wavelength) leading to an avalanche of identical photons. The mechanism which does this is stimulated emission; therefore, what is desired is a physical situation in which the rate of stimulated emission is greater than the rate of absorption or of spontaneous emission. The word laser, which is now accepted as a noun, was originally an acronym for light amplification by stimulated emission of radiation.

The reader can observe the rate equation appears from nowhere and has no justification, but stipulates a new process (stimulated emission) which is non-trivial. This is true, but this has proven, over time to be an accurate model of the world, and so it has been retained. We take the equation above as valid and will examine it for the implications it has for lasing.

Let us now make some observations about the equation above and see what it indicates about a lasing system.

First, it describes dynamic equilibrium. In the material, electrons are constantly absorbing and emitting photons, but the population of excited and ground state electrons and photons stays constant. The units of each of the terms on each side of

the equation are rates (/cm^3 s). When these transition rates are equal, the equation describes a steady-state situation; in thermal equilibrium, the populations can be described by a Boltzmann distribution and the relative size of the populations is as given in Eq. 2.14.

In equilibrium, the population of the higher energy state is always lower than that of the lower energy state, and therefore the rate of absorption is always greater than the rate of stimulated emission: $BN_2N_p(E) > BN_1N_p(E)$ (the absorption rate is always greater than the stimulated emission rate *in thermal equilibrium*). The absorption rate is not only greater, but enormously greater. In a typical semiconductor laser, $E_2 - E_1 \sim 1$ eV, which gives the relative population of ground and excited states as exp (−40) at room temperature. Because in equilibrium $N_2 << N_1$, stimulated emission is much less than absorption, and therefore in equilibrium lasing is not possible.

This means practical *lasing systems must be driven* in some non-equilibrium way, generally either optically or electrically. It is not possible to drive something thermally and achieve a dominant stimulated emission. Practical lasing systems are usually composed of (at least) three levels: an upper and lower level, between which the system relaxes and emits light, and third, pump level, where the system can be excited. This will be illustrated in Sect. 2.6.

In addition, for lasing to occur, the spontaneous emission rate must also be much less than the stimulated emission rate. While both processes produce photons, the spontaneous emission photons are emitted at random times and are thus in random phases compared to the coherent photons generated by stimulated emission. These photons thus do not really contribute to the coherent lasing photons. For a lasing system, $BN_2N_p(E) > AN_2$.

This may or may not be possible depending on the relative values of A and B and various Ns. We note that a *higher photon density*, N_p, certainly makes the balance favorable. There is much more stimulated emission at higher photon density than at lower photon density. Hence, for stimulated emission to dominate, it is beneficial to have a higher photon density. This is achieved in a *laser by always having some cavity mechanism, based on mirrors or other wavelength-selective reflectors, to achieve a high photon density inside the cavity.*

The first equation (stimulated emission greater than absorption) implies that the *lasing system is non-equilibrium* ($N_2 > N_1$) and is called *population inversion*. The second equation (stimulated emission greater than spontaneous emission) implies a high photon density. *These two conditions taken together form a mathematical model for a physical basis for a lasing system.*

$$BN_2N_p(E) > AN_2 \xrightarrow{\text{implies}} \text{high photon density } N_p$$

$$BN_2N_p(E) > BN_1N_p(E) \xrightarrow{\text{implies}} \text{nonequilibrium system with } N_1 < N_2$$

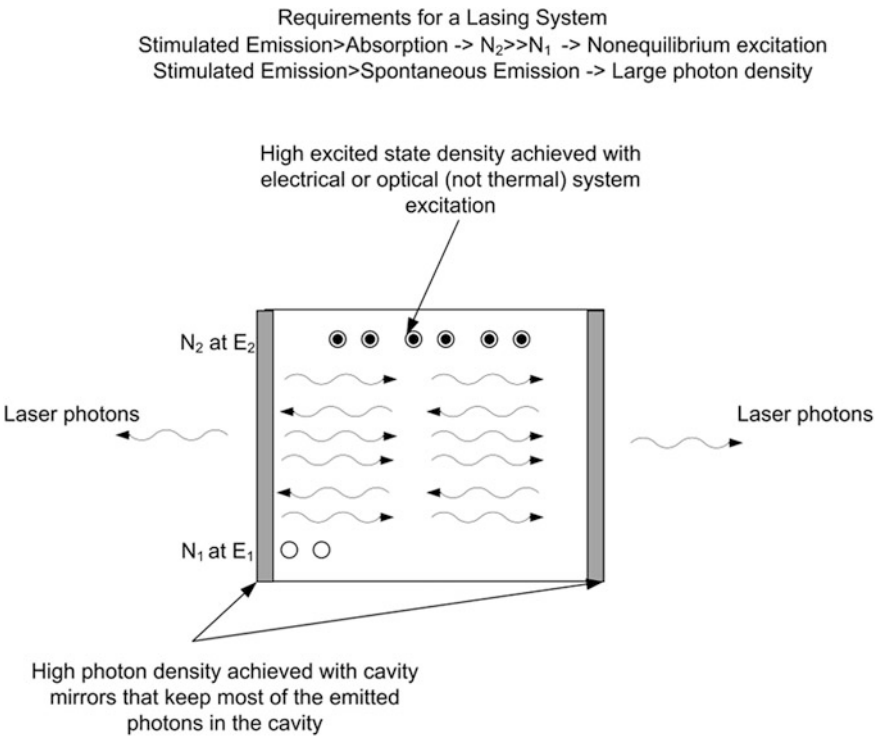

Requirements for a Lasing System
Stimulated Emission>Absorption -> N2>>N1 -> Nonequilibrium excitation
Stimulated Emission>Spontaneous Emission -> Large photon density
High excited state density achieved with electrical or optical (not thermal) system excitation
N2 at E2
Laser photons
Laser photons
N1 at E1
High photon density achieved with cavity mirrors that keep most of the emitted photons in the cavity

Fig. 2.6 Requirements for a lasing system and the way they are implemented in practice. Non-equilibrium pumping is done electrically, or optically, to excite most of the states. A high photon density is achieved by mirrors or other sorts of optical reflectors to maintain a high photon density inside the cavity. A laser usually looks similar to this conceptual picture

The first condition means that we cannot construct a laser that will just heat up and lase. Any heat-driven process is by definition a thermal equilibrium process, and in such processes absorption, rather than emission, will always dominate. This non-equilibrium requirement is realized in real laser systems by having them powered—for example, in semiconductor lasers, the holes and electrons are electrically injected rather than thermally created. These requirements are illustrated in Fig. 2.6. The portion of a lasing system which is in population inversion is called the gain medium.

In the next two sections, we are going to talk about the qualitative differences between spontaneous emission, stimulated emission, and lasing, and give some examples about how these two requirements for lasing systems (driving force and high photon density) are implemented in practice.

2.5 Differences Between Spontaneous Emission, Stimulated Emission, and Lasing

Figure 2.7 illustrates the spectra of some systems dominated by lasing, spontaneous and stimulated emission, to give some intuition to the idea of lasing as a beam of coherent photons and some idea of what is meant by lasing. There is no clean mathematical definition of lasing; the sense of lasing is a monochromatic beam of photons that is dominated by stimulated emission. Figure 2.7 shows the spectrum for a standard semiconductor laser (a distributed feedback laser) whose spectra are dominated by stimulated emission and shows a near-monochromatic one wavelength peak; the spectrum of a light-emitting diode, whose emission shows a broad peak characteristic of spontaneous emission from the band gap of the semiconductor; and finally, a doped Eu system which has achieved population inversion but not an extremely high photon density and as such exhibits a spectral narrowing but not to the extent seen in (a). We will refer back to this figure and discuss some of the details of the spectra later in this book; for now, we just wish the reader to note that one laser characteristic is an extremely narrow spectra, and that there is a different qualitative character to each of the different mechanisms of stimulated emission, spontaneous emission, and lasing.

In the middle figure, also note that the power density where the system starts to exhibit substantial stimulated emission ($BN_2N_p > AN_2$) is quite clear. There is also a dynamic element in these lasing systems. Because the population must be inverted ($N_2 > N_1$), the amount of time an excited state exists before it relaxes is extremely important and can influence properties like the threshold of lasing systems. This also will be talked about in greater detail later.

We note also that absorption can be considered a 'stimulated' process, which is the opposite of stimulated emission.

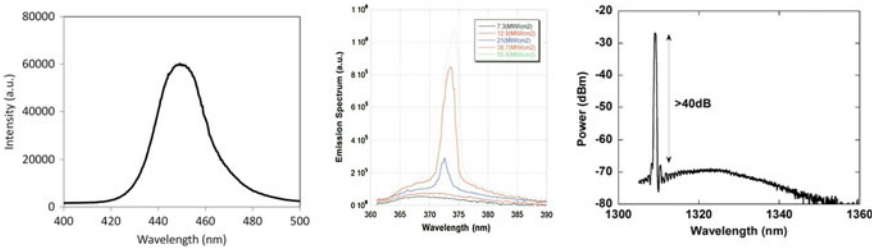


Fig. 2.7 Spectra of some semiconductor-based light-emitting systems. Left, some light-emitting diode spectra with bandwidth of 40–50 nm; center, spectra of a doped Eu system which is showing substantial stimulated emission (a positive feedback cascade of photons at peak wavelength, with a bandwidth of a few nm) but not lasing; right, finally, a full single-mode distributed feedback laser, showing very narrow linewidth (<1 nm)

2.6 Some Example of Laser Systems

All lasers consist of a gain medium, a method of non-equilibrium pumping, and a cavity defined by mirrors or another mechanism to obtain a high photon density. We now show specific examples illustrating how these three properties are achieved. Because the bulk of the book will discuss semiconductor lasers, these examples are going to be taken from other laser systems.

Apart from the gain medium, this will also show the various ways in which optical cavities are formed to contain the photons.

First, an *Er-doped fiber laser* has the atomic levels of the erbium (Er) atom as the gain medium, optical pumping as the means for inducing non-equilibrium, and a Bragg grating cavity integrated into the fiber as the cavity mirror to achieve a high photon density.

Second, we will talk about a common red HeNe gas laser, which has the Ne atomic levels as the gain medium, a high-voltage AC source as the method of electrically exciting (pumping) the molecules, and high-reflectivity mirrors defining the cavity.

2.6.1 Erbium-Doped Fiber Laser

As an illustration, Fig. 2.8 depicts the energy levels and the physical structure of an erbium (Er)-doped fiber laser. This structure is similar to an Er-doped fiber amplifier, but with an engineered cavity. An optical fiber is fabricated doped with optically active Er atoms, and a simplified version of the Er atomic energy level is shown at above left. Pump light at 1 μm excites the atoms into an excited state (the $4I_{15/2}$ state), which then rapidly ($\sim$ns) relaxes into a state with a band gap at around 1 μm (the $4I_{13/2}$ state). This state has a lifetime of $\sim$ms, and so the system can be put into population inversion in which the density of atoms in the $4I_{11/2}$ state is much higher than the $4I_{13/2}$ state. Here, the three states ($4I_{15/2}$, $4I_{13/2}$, and $4I_{11/2}$) are the pump level, upper level, and lower level, respectively.

The *dynamics* are actually critical to this system. If the relaxation between $4I_{15/2}$ and $4I_{13/2}$ was slower, or the relaxation between $4I_{13/2}$ and $4I_{11/2}$ was faster, it would be much harder to achieve 'population inversion' system in which the population of $4I_{13/2} > 4I_{11/2}$, as required for lasing.

The other requirement for lasing is high photon density. This is accomplished by the Bragg gratings integrated into the fibers, which confine most of the 1.55 μm photons into the fiber laser cavity. In order to allow the pump light in freely, these gratings have to have a low reflectivity at 1 μm. This system produces a device which, when high-intensity 1 μm light is coupled into the fiber, produces a monochromatic beam of 1.55 μm light out.

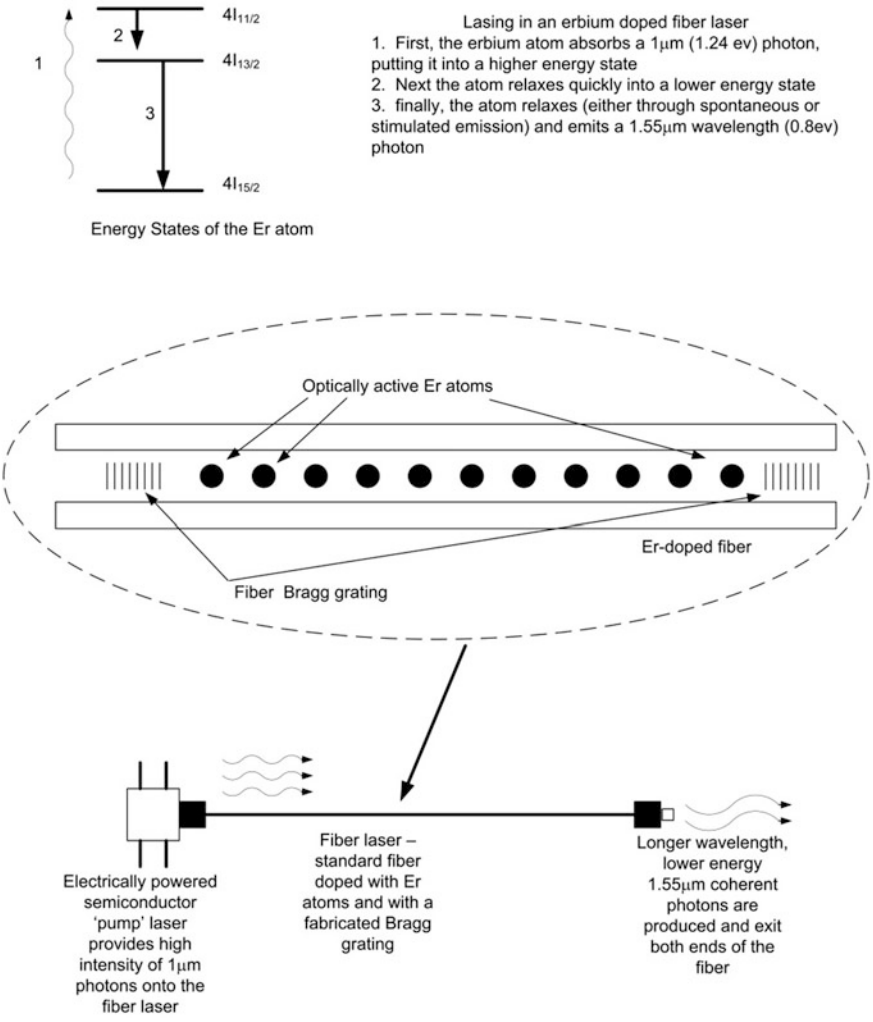

1
2
3
$4I_{11/2}$
$4I_{13/2}$
$4I_{15/2}$
Energy States of the Er atom
Lasing in an erbium doped fiber laser
1. First, the erbium atom absorbs a 1μm (1.24 ev) photon, putting it into a higher energy state
2. Next the atom relaxes quickly into a lower energy state
3. finally, the atom relaxes (either through spontaneous or stimulated emission) and emits a 1.55μm wavelength (0.8ev) photon
Optically active Er atoms
Er-doped fiber
Fiber Bragg grating
Electrically powered semiconductor 'pump' laser provides high intensity of 1μm photons onto the fiber laser
Fiber laser – standard fiber doped with Er atoms and with a fabricated Bragg grating
Longer wavelength, lower energy 1.55μm coherent photons are produced and exit both ends of the fiber

Fig. 2.8 An erbium-doped fiber laser. As shown, population inversion is achieved between the $4I_{13/2}$ and $4I_{11/2}$ levels by optical pumping, a non-equilibrium process. High photon density is achieved by Bragg mirrors, which keep most of the 1.55 μm photons in the laser length of the fiber

2.6.2 HeNe Gas Laser

The traditional red laser that is often used in optics laboratories is a HeNe gas laser. The schematic picture of such a laser and its mechanism for operation is shown in Fig. 2.9. The gain medium is the HeNe molecules that are sealed in the tube. A high DC voltage is applied which creates electrons which excite a He atom. The He atom then transfers its energy to a Ne atom. The Ne atom then relaxes by radiative stimulated emission to a lower level, emitting a red photon at $\lambda = 632$ nm in the

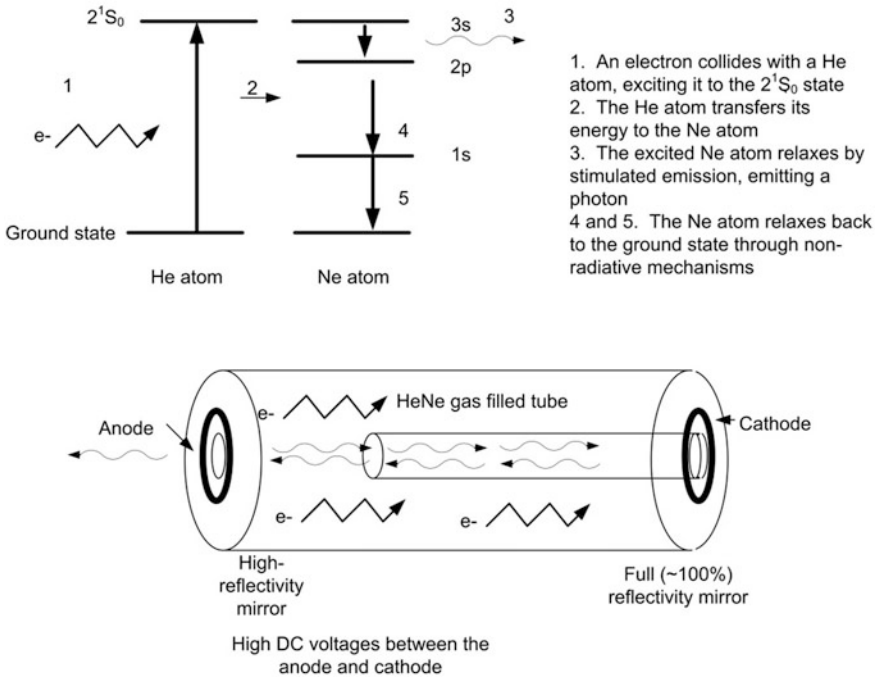


Fig. 2.9 A HeNe gas laser, showing the gain medium (the Ne atom), the high photon density (created by high-reflectivity mirrors), and the method for non-equilibrium pumping by electronic excitation. The bottom shows the physical picture of a HeNe laser; the tube is the active laser region, while the area around it is a reserve gas cavity

process. Even though the light has already been emitted, the Ne atom then has to relax through several more levels non-radiatively down to the ground state to be reused. Finally, the photons are kept in the cavity by the mirrors at each end of the tube. The reflectivity is typically $\sim 99\%$ or more, so the photon density inside the laser is much higher than the photon density right outside the cavity.

There are several atomic levels to the Ne atom. By tailoring the cavity to confine photons of different wavelengths (a mirror specific to red, green, or infrared wavelengths), the same system can be induced to lase in the green or infrared as well as red. Commercial HeNe lasers at all these wavelengths can be purchased.

In Fig. 2.9, the upper portion shows the atomic-level picture of the mechanism for operation of the HeNe laser. The molecule is initially excited, and the relaxation time from the excited state is long enough that the system can be put into population inversion. Once population inversion is achieved, lasing occurs because stimulated emission dominates and the photon density is kept high with the highly reflective facets. The laser cavity is shown at the bottom.

Semiconductor lasers will be covered extensively in the following chapters. In general, they have electrical injection as the pumping method, with the conduction and valence bands serving as the gain medium. There are many mirror methods available in semiconductor lasers; the simplest one is simply the mirror formed when the semiconductor with the refractive index $n = 3.5$ is cleaved, and an interface with the air ($n = 1$) is formed.

2.7 Summary and Learning Points

A. *Distribution functions* describe the probability that an existing energy state is occupied. They describe systems in thermodynamic equilibrium. Different functions are appropriate to different situations. The Fermi–Dirac distribution function is applicable to particles which follow the exclusion principle (electrons or holes); the Bose–Einstein is applicable to photons or protons or other particles who like to aggregate; the Boltzmann distribution function is the classical approximation to both.
B. The *density of state* function is the number of states at a given energy in a system. The density of photon states in a black body can be calculated and that, combined with the appropriate distribution function, gives the black body emission spectra.
C. By equating the rates of particle relaxation and excitation (in a 'dynamic' equilibrium), the same picture of black body emission spectra can be obtained (provided that the two Einstein B coefficients are equal). This model resulted in defining the (new) mechanism of light emission called *stimulated emission*, in which a photon impinges on an excited atom and causes it to emit another photon of the same wavelength and phase. It is this mechanism that is responsible for lasing.
D. A *laser* is a coherent light source generated by stimulated emission. Hence, stimulated emission has to dominate over both absorption and spontaneous emission. These criteria require a lasing system to:

    i. be in population inversion, with more of the gain medium in the excited state than in the ground state;
    ii. have a high photon density N_p, which requires mirrors or facets to surround the lasing system.

E. Because of the population inversion requirement, a laser cannot be driven thermally. Lasers are non-equilibrium systems.

2.8 Questions

Q2.1. Define stimulated emission of radiation.
Q2.2. Explain how the temperature can be measured from a black body spectrum.
Q2.3. Explain in your own words the statistical thermodynamics perspective of black body radiation.
Q2.4. Explain in your own words the microscopic view of black body radiation.
Q2.5. Define the term 'distribution function.'
Q2.6. Define the term 'population inversion.'
Q2.7. What distribution function is appropriate for photons? For electrons?
Q2.8. When is it appropriate to use the Gaussian distribution function?
Q2.9. Define the term 'density of states.'
Q2.10. If the k-value of a particular photon state is very large, is the wavelength of that photon high or low? Is the energy of that photon high or low?
Q2.11. List the three requirements for any lasing system.
Q2.12. Explain how these requirements are met in your own words for the two types of lasers discussed in the chapter.
Q2.13. What are the three levels in the HeNe laser system?

2.9 Problems

P2.1. Show that Eq. 2.11 reduces to Plank's expression for a black body spectrum, Eq. 2.1.
P2.2. Show that for a system in thermal equilibrium, the coefficient of stimulated emission B_{21} is equal to the coefficient of stimulated absorption B_{12}. (Hint: Use the fact that the $N_2/N_1 = \exp(-\Delta E/kT)$, and the fact the Einstein and Plank black body spectra must agree).
P2.3. A photon has a wavelength of 500 nm.

(a) What color is it?
(b) What is its energy, in?

a. J
b. eV.

(c) What is the magnitude of its spatial propagation vector k?
(d) Find its frequency in Hz.

P2.4. (This problem is given by Kasap,[2] and used by permission). Given a 1 μm cubic cavity, with a medium refractive index $n = 1$:

(a) Show that the two lowest frequencies which can exist are 260 and 367 THz.
(b) Consider a single excited atom in the absence of photons. Let p_{sp1} be the probability that the atom spontaneously emits a photon into the (2, 1, 1) mode, and p_{sp2} be the probability density that the atom spontaneously emits a photon with frequency of 367 THZ. Find p_{sp2}/p_{sp1}.

P2.5. This problem explores the influence of dynamics on the populations of the erbium atom levels. In Fig. 2.8, the energy levels of the erbium atom are pictured.

(a) If a population of Er atoms absorbs 10^{18} photons/second, but the lifetime of the excited state is 1 ns, what is the steady-state population of atoms in the $4I_{11/2}$ state?
(b) If the lifetime of the $4I_{13/2}$ state is 1 mS, what is the steady-state population of the $4I_{13/2}$ state?
(c) How many 1.55 μm photons are emitted per second?

[2]S. O. Kasap, *Optoelectronics and Photonics: Principles and Practices*. Upper Saddle River, NJ: Prentice Hall, 2001.

Check for updates

Semiconductors as Laser Materials 1: Fundamentals

3

You can observe a lot by just watching.
—Yogi Berra

Abstract

The descriptive overview provided in this chapter is a prelude to the mathematical modeling of semiconductor and optical properties that follows in later chapters. Here, we discuss the relevant properties of semiconductor quantum wells from the point of view of applications for semiconductor lasers. First, we introduce the general idea that semiconductor lasers are composed of mixtures of semiconductors designed to select the appropriate lattice constant and band gap. The physical limits of mixing of different semiconductors are covered. Practical factors that influence the use and fabrication of semiconductors for lasers including factors such as direct and indirect band gaps, and strain and critical thickness, are discussed.

3.1 Introduction

As seen in Chap. 2, lasers can be constructed with many different material system, and different lasers have different applications. For example, HeNe lasers are used as coherent sources for optical experiments. High-power Ti: Sapphire lasers can be used to generate very short, high-intensity optical power bursts, and CO_2 gas lasers can produce extremely high-power bursts that can be used to machine materials. This textbook focuses on the semiconductor lasers used in optical communications.

© Springer Nature Switzerland AG 2020

D. J. Klotzkin, *Introduction to Semiconductor Lasers for Optical Communications*,
https://doi.org/10.1007/978-3-030-24501-6_3

In this chapter, we discuss the basics of semiconductors as a lasing medium and the practical details of designing and making these complex laser heterostructures. First, we address the details of designing heterostructures of different compounds, and we cover considerations of growing thin films of these heterostructures. Finally, we discuss the band structure of real semiconductors.

3.2 Energy Bands and Radiative Recombination

The semiconductor is the gain medium in a semiconductor laser. A very simple diagram of the electron structure of a semiconductor is shown in Fig. 3.1. In general, a semiconductor has a valence band, in which (effectively) holes (positive

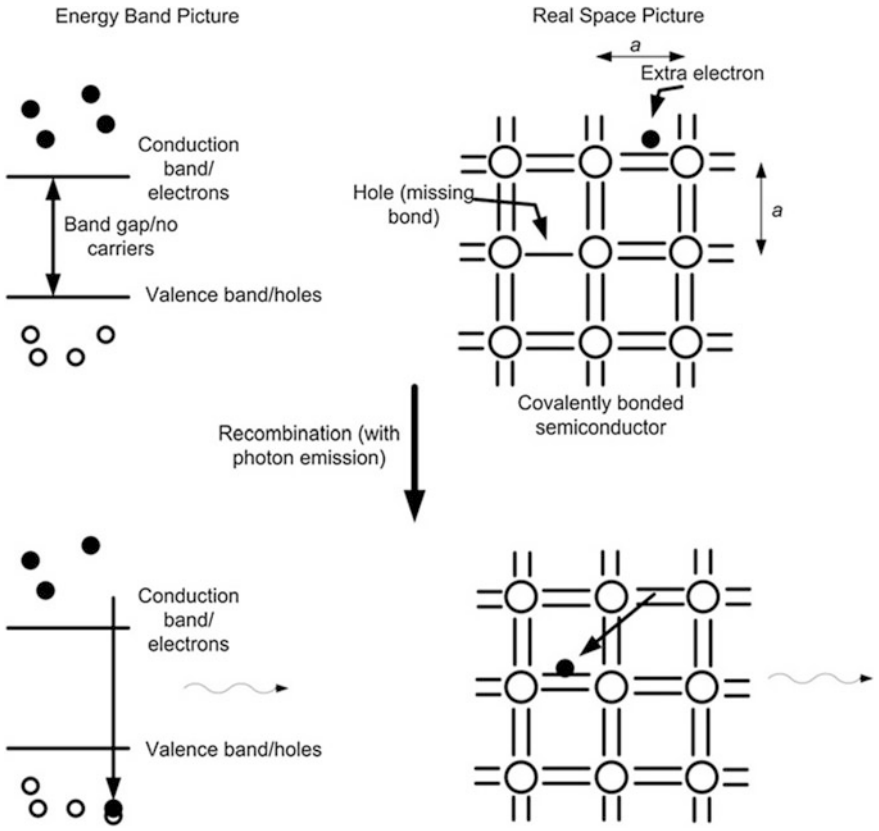


Fig. 3.1 Basics of semiconductors for laser application. They emit light due to recombination of electrons and holes across the band gap. The distance *a* in (c) and (d) represents the lattice constant of the semiconductor

charges) exist and conduct current, and a conduction (or electronic) band in which electrons (negative charges) exist and conduct current.

Usually, semiconductors are doped to influence their electrical properties. Doping means that the semiconductor (say Si, for example) has some amounts of other atoms incorporated into it (say, B). Here, B has only three electrons per atom in its outer shell, so the doped semiconductor has an average of slightly less than four electrons/atom. These missing electrons in the valence band act as conductors. In doped semiconductors, one or the other of these charge carriers dominates: in the example, here, the charge conductors would be holes with a positive sign.

Because of the periodicity of the crystalline array, the energy *levels* associated with an *atom* become the energy *bands* within a *crystal*. These leave a band gap of forbidden electron energies. In semiconductor compounds, the average of four electrons per atom is precisely enough to fill up the lowest energy level and leave the higher energy levels empty. This situation creates the useful semiconductor property of a moderate band gap, and conductivity that is easily controlled by doping.

Real semiconductor bands are much more complicated than the description implied by the single band gap number. For example, only some semiconductors—those with what are called direct band gaps, like GaAs and InP—support electron-hole recombinations that emit light. These and other qualitative details of the bands will be discussed at the end of the chapter.

In the context of lasers, we are more concerned with electron and hole recombination rather than with conduction. When an electron recombines with a hole, eliminating them both, the resulting energy can be emitted in the form of a photon through radiative recombination. Hence, the band gap (the difference in energy between the hole and electronic levels) determines the value of the wavelength of light emitted by a particular semiconductor.

Figure 3.1 shows the process from both an energy diagram view and a physical 'real space' view. A photon is emitted when an electron in the conduction band recombines with a hole in the valence band, eliminating both.

In general, the more readily a material recombines and emits light spontaneously (spontaneous emission), the better the material works as a laser (with stimulated emission). The Einstein model of stimulated/spontaneous emission predicts a relationship between the A and B coefficients of spontaneous and stimulated emission, and in practice a good light emitter (like a direct band gap semiconductor) works well either in spontaneous emission, as a light-emitting diode, or with stimulated emission in a laser configuration, with mirrors and a mechanism for non-equilibrium pumping.

In telecommunications lasers, the band gap largely determines the wavelength of light emitted from the semiconductor. But how is the band gap determined? We will discuss the answer to the question in later sections.

3.3 Semiconductor Laser Material System

For semiconductor laser applications, we need a material with a particular band gap that emits light at a particular wavelength. Typically, the material is grown on a semiconductor substrate (for example, a laser may be made from InGaAs quantum wells on a GaAs substrate). The lattice constant of the material (which is the characteristic size of the unit cell, '*a*', as illustrated (in 2D) in Fig. 3.1, and (in 3D) in Fig. 3.3) has to closely match the lattice constant substrate for a successful growth. To obtain a working laser, the material has to be nearly *lattice-matched* to the substrate on which it is grown and have the right bandgap for a particular wavelength of light.

(As an aside, laser material sometimes intentionally is not perfectly lattice matched—it is designed for it to be slightly different than that of the substrate. We defer discussion of this topic until later in the chapter).

As a concrete example, let us talk about the InGaAsP laser system, commonly grown on InP and used across the important telecommunications spectrum from 1.3 to 1.6 μm. The system has in it four binaries (fundamental III-V compounds made of two elements, like GaAs or InP), whose band gap and lattice constant are listed in Table 3.1.

The layers from which communications lasers are made are the quantum well layers and are usually grown on an InP substrate. Whatever the wavelength, it is important that the value of lattice constant of the layer be close to 5.8686 Å. The utility of this material system stems from the ability to grow nearly perfect heterostructures of the four basic elements, with In and Ga freely interchangeable and As and P freely interchangeable. The quaternary compounds of $In_xGa_{1-x}As_yP_{1-y}$ can span a broad array of band gaps and lattice constants.

Figure 3.2 shows the bandgap and lattice constant of the binaries in Table 3.1 (and many others) plotted on a graph with band gap (or emission wavelength) shown on the *y*-axis and lattice constant is shown on the *y*-axis. To grow a 1.55 μm laser lattice-matched to InP (a very common case) the composition should lie at the intersection of the line $y = 1.55$ μm and $x = 5.8686$ Å. The intersection of the two constraints lies somewhere within the parameters spanned by the four binaries, suggesting that there is some compound of InGaAsP (denoted by $In_xGa_{1-x}As_yP_{1-y}$) that will match both lattice constant and desired band gap.

Table 3.1 Band gap (eV) and lattice constant (Å) of binaries in the InGaAsP family

Binary	Band gap (eV)	Lattice constant (Å)
InP	1.34	5.8686
InAs	0.36	6.05838
GaAs	1.43	5.65315
GaP	2.26	5.4512

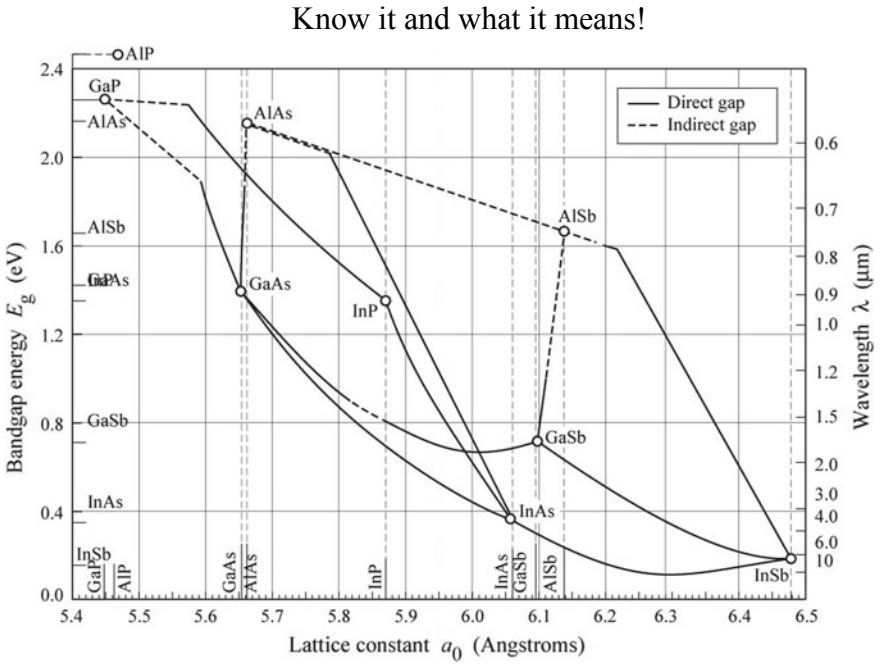


Fig. 3.2 Semiconductor chart showing properties (lattice constant and band gap, in both eV and μm) versus composition. The lines between pairs of binary semiconductors represent the properties of heterostructures of those two binaries (a ternary). Quaternary compounds can access all of the area bounded by their four boundaries. From E. F. Schubert, Light Emitting Diodes, Cambridge University Press, 2006, used by permission

What Si is to ordinary CMOS electronics, III-V compound semiconductors are to telecommunications optoelectronics. The utility of InP-based lasers for telecommunications applications arises from the fact that its bandgap overlaps both 1.55 and 1.3 μm, which are the low loss and low dispersion windows for optical fiber, respectively. In a different role, GaAs-based lasers are used for as a key amplifier component to make lasers around 1 μm wavelength.

To give a physical picture of the semiconductor lattice, Fig. 3.3 shows the zinc blende lattice of both GaAs- and InP-based heterostructures (in fact, it is the same structure for Si lattices also, just with only Si atoms throughout). The length of the unit cell is the lattice constant a. The white dots are Group III atoms, and the black dots are Group V atoms. In this lattice, any Group III atom can occupy any Group III site. Each Group III atom (with valence III) is surrounded by four Group V atoms of valence 5, so the structure as a whole (undoped) has an average valence of 4.

In doped semiconductors, the dopant atoms occupy some of the positions formerly occupied by Group III or Group V atoms. In that case, the crystal is still perfect, but has a shortage or excess of electrons over its nominal number of four electrons/atom.

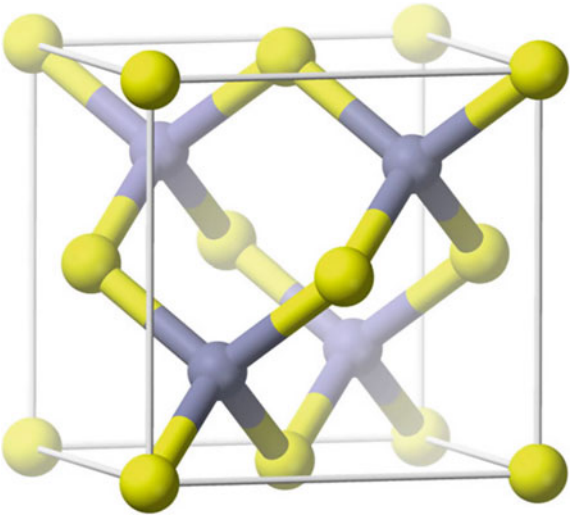


Fig. 3.3 A picture of the zinc blende lattice, showing each group III (Ga) atom surrounded by 4 group V (As) atoms, and each group V atom surrounded by four group III atoms. Any group III atom can occupy any group III site, and by variations of the composition, the bandgap lattice constant, and other associated properties can be picked. From Wikipedia, http://en.wikipedia.org/wiki/Zincblende_%28crystal_structure%29#Zincblende_structure,current 9/1/2013

3.4 Determining the Band Gap

If we are constrained by nature to use only binary compounds with fixed bandgap, we would not have semiconductor laser-based optical communications. There simply are not enough wavelengths! However, we can mix and match atoms to achieve materials with a wide range of band gaps and wavelengths.

The wavelength λ at which a material with a given bandgap E_g emits is given by

$$\lambda = \frac{hc}{E_g} \tag{3.1}$$

which comes from Plank's relation between the energy and wavelength λ of the photon. The easy way to remember this is the constant hc = 1.24 eV-μm. So, the equation above can be written as

$$\lambda(\mu\text{m}) = \frac{1.24\ \text{eV-}\mu\text{m}}{E_g(\text{eV})} \tag{3.2}$$

which means that if the band gap is given in eV (the usual unit of band gaps), dividing 1.24 by that number will give the wavelength in μm.

Example: What bandgap semiconductor is necessary to emit a very long wavelength 10 μm photon? How does that compare to the thermal energy kT at room temperature?

Solution: If the hypothetical semiconductor emits at 10 μm, the bandgap (in eV) can be determined to be 1.24 eV-μm/10 μm = 0.12 eV. The thermal energy kT at room temperature is 0.026 eV, about ¼ of this bandgap. This device would probably only work at very low temperatures.

3.4.1 Vegard's Law: Ternary Compounds

Let us now demonstrate how we can design a heterostructure with a particular bandgap. This is easiest illustrated by an example, given below and based on a ternary compound.

Example: What mole fraction x of In in $In_xGa_{1-x}As$ will result in a material that emits light at 1 μm wavelength?

Solution: The compound $In_xGa_{1-x}As$ is made up of GaAs and InAs. We assume that the bandgap property is a linear interpolation of the bandgaps of GaAs and InAs. The energy corresponding to 1 μm light emission is 1.24 eV-μm/1 μm or 1.24 eV, so that the desired bandgap is 1.24 eV at room temperature. Using the data from Table 3.1, the equation
$1.24\ \text{eV} = xE_g(\text{InAs}) + (1-x)E_g(\text{GaAs}) = x0.36 + (1-x)(1.43)$ gives $x = 0.17$. Thus, a mole fraction of In of $x = 0.17$ will give a material with a bandgap of 1.24 eV.

Let us look at another example calculating the property of an existing semiconductor.

Example: What will the lattice constant be of $In_{0.17}Ga_{0.83}As$?

Solution: In the same way that energy gaps average, lattices constants average. In this case, the lattice constant a of $In_{0.17}Ga_{0.83}As$ will be $0.83a(InAs) + 0.17a(GaAs) = 5.7222$ Å, where a(compound) represents the lattice constant of that compound.

Notice that of course the total number of Group III and Group V atoms are the same, since semiconductors have equal numbers of Group III and Group V atoms; for example the compound $In_{0.2}Ga_{0.1}As$, which has more Group V than Group III atoms, is certainly not a semiconductor and in all likelihood could not be fabricated at all.

This linear interpolation between binary compounds is called Vegard's law and serves as a very useful first approximation for how we design material composition for a given bandgap and lattice constant. In general, for a property Q of a ternary alloy $A_{1-x}B_xC$,

$$Q(A_{1-x}B_xC) = (1-x)Q(AC) + xQ(AB) \tag{3.3}$$

where $Q(AC)$ and $Q(AD)$ are the properties of the associate binaries, In practice, what is usually done is to approximate the composition for a particular bandgap by some kind of estimation technique, such as this one. Then the material is grown, and the composition is measured. The small variations in the composition are corrected in subsequent growths. (How the material is grown is discussed in Sect. 3.5.1, upcoming, and in Chap. 10).

From Fig. 3.2, a linear interpolation is perfectly appropriate to approximate the properties of $In_{1-x}Ga_xAs$. By adjusting the composition of the heterogenous semiconductor, *the bandgap, refractive index, and lattice constant can be selected*. The power and the utility of these compounds are the ability to engineer properties (such as bandgap, refractive index, and lattice constant) to whatever is required by mixing together Group III and Group V atoms. Ternary compounds (such as $In_{1-x}Ga_xAs$) have one degree of freedom (the fraction of Ga atoms) and so by picking a lattice constant, the band gap is specified. Quaternary compounds (like $In_{1-x}Ga_xAs_{1-y}P_y$) have two degrees of freedom, and so (within certain limits) can independently pick both bandgap and lattice constant. This freedom allows for design of layers that can be grown on InP with the desired strain and bandgap.

A broad range of materials with different bandgaps (or wavelengths) can be made by making heterostructures or combinations of binary compounds. This averaging process consists of randomly arranging group different Group III atoms on Group III sites, and Group V atoms on Group V sites as pictured in Fig. 3.3. The whole compound is always constrained to having equal number of group III and group V atoms.

3.4.2 Vegard's Law: Quaternary Compounds

Please look again at Fig. 3.2 shows the bandgap and lattice constant of the four binaries. Bounded by the four binaries of Table 3.1, it is apparent that a range of bandgaps (from 0.36 eV of InAs to 2.3 eV for GaP) can be achieved on a range of lattice constants from 5.45 to 6.05 Å, and in particular lattice-matched to InP (5.86 Å). How does the parameter (lattice constant, band gap, or effective index) depend on composition for these quaternaries?

The basic result, which we will present here, is that for the quaternary $A_{1-x}B_xC_yD_{1-y}$ the property $Q(A_{1-x}B_xC_yD_{1-y})$ is given by

$$Q(x,y) = xyQ(BC) + x(1-y)Q(BD)] + (1-x)(y)Q(AC) + (1-x)(1-y)Q(AD) \tag{3.4}$$

This formula gives a good start to get a fixed bandwidth, based on the assumption of perfect linear interpolations between the binaries. While this formula gives a good first-order approximation, usually slight refinements of composition are necessary to obtain the exact desired property. A careful look at Fig. 3.2 shows that dependence of properties on composition is rarely exactly linear.

3.5 Lattice Constant, Strain, and Critical Thickness

Now that we have discussed growing a material with given properties like bandgap, let us focus in this section on the growth of thin films on a substrate. Thin films are important because the vast majority of lasers are made by depositing thin films on a substrate to form quantum wells. Hence, what happens when thin films are deposited on a substrate—both to their electronic properties and physically—is extremely important.

The lattice constant is the fundamental size of the unit of a semiconductor. A mismatch in lattice constant between the thin film and the material it is being grown on (the substrate) causes *strain* in the material. Just like a spring when it is compressed or stretched, is strained and exerts force to return to its desired dimension, a layer of material

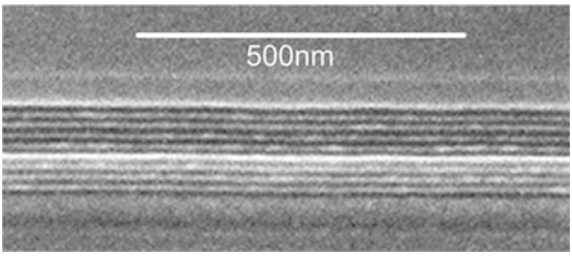


Fig. 3.4 An SEM of a semiconductor quantum well structure. The active region consists of quantum wells surrounded by barrier layers, with the entire stack less than 1400 Å total. The thin films have to match the lattice constant of the substrate within a few percent

deposited on a material of different lattice constant also is strained. A strained layer cannot be grown indefinitely—when it is grown too thick, the atomic bonds will break (or the springs will pop back to their normal size), creating dislocations, or missing atomic bonds. The maximum thickness a strained layer can be grown without incurring dislocations is called its *critical thickness* and depends on the degree of lattice mismatch in the material. When growing these thin layers which are used in lasers, strain and critical thickness are very important, because it is imperative to good laser performance to have a low defect density. Dislocations resulting from strain are a kind of material defect. Figure 3.4 shows some of the thin layers forming the quantum wells that define the laser active region.

3.5.1 Thin Film Epitaxial Growth

For these devices to emit light, they have to be assembled from nearly perfect crystals. Imperfections, like missing atoms or extra atoms, create recombination centers which cause carriers to recombine and create heat, rather than light. This engineering requirement that semiconductor lasers be nearly perfect crystals is part of the reason that fabrication of semiconductor lasers is half science and half engineering (with the growth of them being half art!). However, it also imposes a specific requirement on the lattice constant of these layers. For devices to work as emitters, these semiconductors thin films need to match, quite closely, the lattice constant of the substrate.

The active semiconductor layers are grown on a semiconductor wafer, called a substrate (InP is a typical substrate). All of the various methods for semiconductor growth (molecular beam oxide, MBE, or metallorganic chemical vapor deposition, MOCVD) deposit atoms onto the existing substrate, with the atoms bonding one-by-one, atomically, to the existing layers.

Let us examine what happens when a layer of material that is not quite the same lattice constant is deposited.

One analogy is stacking foam bricks of one size on a wall of bricks of a different size. If the size of the bricks being stacked is only slightly different than that the bricks already on the wall, then the new bricks can be squeezed or stretched slightly but fit in, matched brick-by-brick, to the bricks already in the wall. This is called strain which is induced in the new layer.

If the new bricks, or new material, are much larger than the substrate, then it is impossible to line up brick-by-brick; nature's solution is then to leave a brick (or a bond) out, and henceforth, match up the new bricks properly. This omitted brick, or atom, is called a dislocation. These dislocations (missing or extra atoms) are fatal for lasers; they act as non-radiative recombination sites, which compete with radiative recombination to consume carriers. Figure 3.5 shows both strain and dislocation.

Quantitatively, the strain f in a thin film is given by the difference in lattice constants between the substrate $a_{\text{substrate}}$ and the film a_{film} as

$$f = \frac{a_{\text{film}} - a_{\text{substrate}}}{a_{\text{substrate}}}. \tag{3.5}$$

The strain f is typically reported as percentage. If the film material lattice constant is larger than the substrate, the film is said to be compressively strained; otherwise, it is said to be tensile strained.

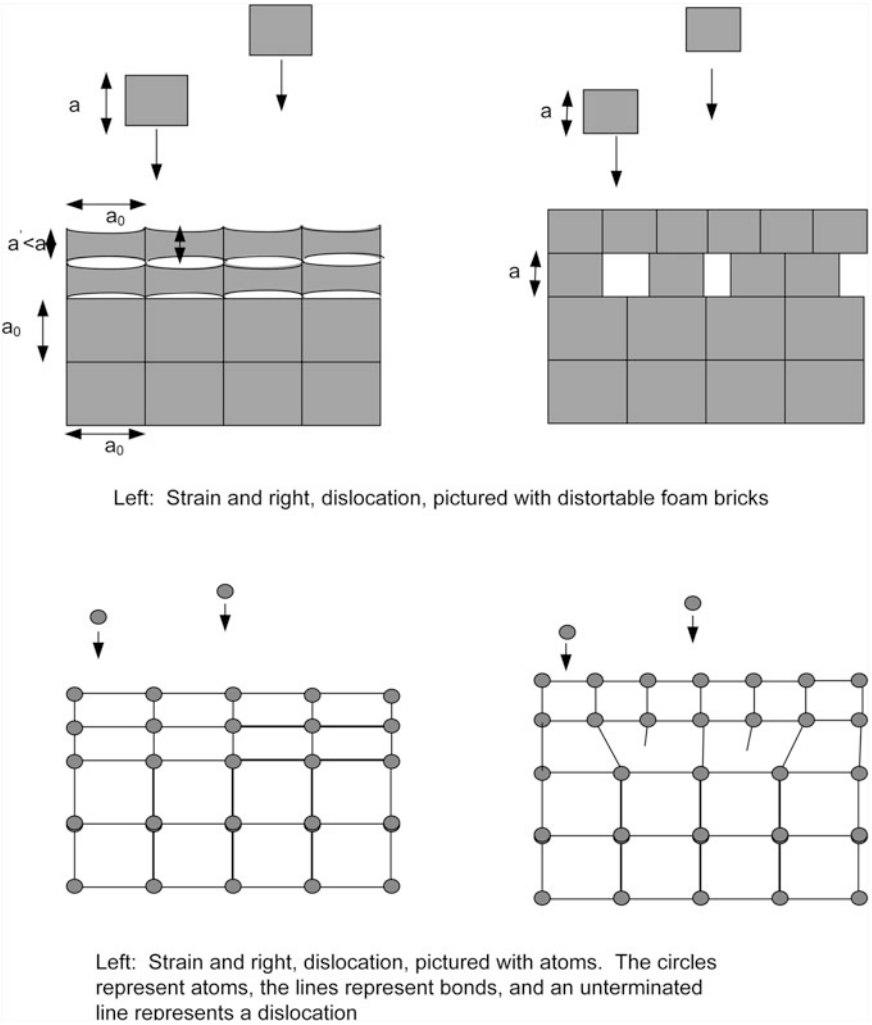

a
a0
a'<a
a0
a0
a
a
Left: Strain and right, dislocation, pictured with distortable foam bricks
Left: Strain and right, dislocation, pictured with atoms. The circles represent atoms, the lines represent bonds, and an unterminated line represents a dislocation

Fig. 3.5 Strain and dislocation. The left side shows that strain results in a distortion (stress) distributed on each of the unit cells (or foam bricks) deposited. On the right, dislocations suffer some energy penalty from missing bonds at the interface but thereafter are perfect crystals. These dislocations at the interface act as non-radiative recombination sites and are deleterious to lasers

Typically, layers have can strains up to about 1% or a little more. A modest amount of strain can be beneficial in improving the speed or other properties of the device, as we will discuss in later chapters.

3.5.2 Strain and Critical Thickness

As one can imagine, the more atomic layers (or springs, or bricks) that are stacked together, the more energy it takes to hold them squeezed into their non-equilibrium

shape. These thin layers can only be grown up to a certain thickness before dislocations start to appear. This thickness is called the critical thickness and is of great important to lasers. Quantum well lasers are made up of quantum wells, which are thin ($\sim$100 Å) layers of one material sandwiched between other, thin layers of material. These layers are usually not quite lattice-matched to their substrate, and so it is important to be aware of the strain and the material limits on how thick these layers can stack up.

One way to envision this is to imagine that nature will pick the lowest energy solution. If there only a few atoms in a thin layer, they will be strained, and match up to the substrate; if there are a large number of atoms in a thick strained layer, it is energetically favorable to have a few broken bonds in one layer, and thereafter grow a relaxed layer with its equilibrium lattice constant not matched to the substrate.

This model of critical thickness, which is based on comparison of dislocation energy and strain energy, is based on the thermodynamic equilibrium of minimum energy. In reality, the layers do not know how thick they will be when they are initially grown. Starting with a few strained layers already, there is a kinetic barrier to switching to a dislocation after fifty or a hundred layers of atoms have been grown. Because of this, layers substantially thicker than the critical thickness can usually be grown without dislocation in practice. But a lot depends on how (deposition rate, and deposition temperature) the layers are deposited.

There are several models of how thick these layers can be (the critical thickness t_c), based on the degree of strain f, and the lattice constant a. The simplest is

$$t_c = \frac{a_{\text{film}}}{2f} \tag{3.6}$$

For example, an InGaAs layer with a lattice constant of 5.67 Å grown on a GaAs substrate with a lattice constant of 5.65 Å would have a compressive strain of 0.35%, and a critical thickness of 800 Å. Such numbers are typical for critical thickness dimensions.

This strain is cumulative, so alternating layers of GaAs and InGaAs on a GaAs will allow a *total* of 800 Å of InGaAs to be grown. However, there is also a strategy used in quantum wells to allow as many different thin layers to be grown as desired. Strain compensation (used in multiple quantum well lasers) pairs compressively strained layers with tensilely strained layers. The net effect is that the strain cancels and very thick layers can be grown. Figure 3.6 shows a typical laser set of quantum wells and barriers, with and without strain compensation.

Example: What is the critical thickness of a layer of $In_{0.17}Ga_{0.83}$ As grown on a GaAs substrate?

Solution: As we see from the previous example, the lattice constant a of $In_{0.17}Ga_{0.83}As$ is 5.7222 Å. Hence the strain is $(5.6532 - 5.7222)/5.6532$, which is compressive,

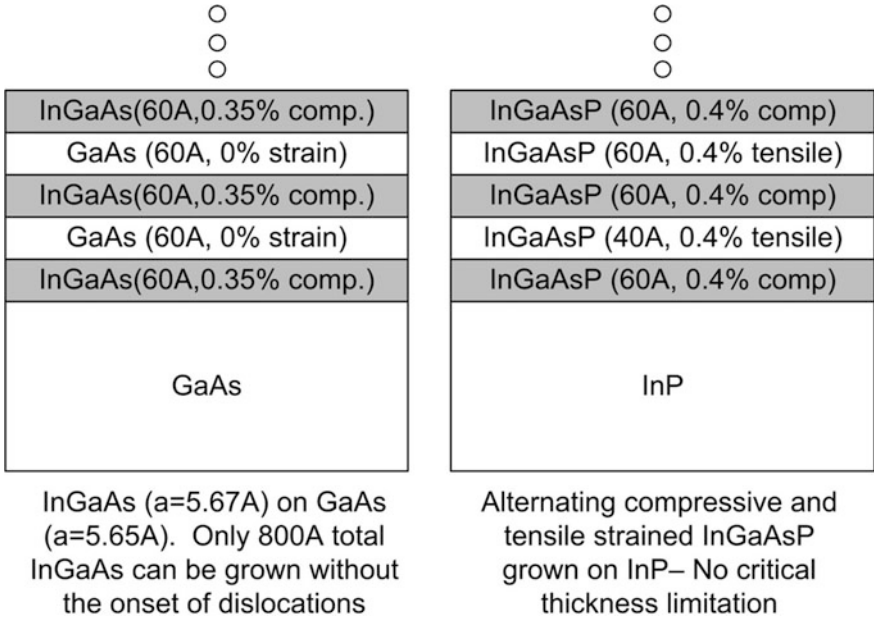

InGaAs(60A,0.35% comp.)
GaAs (60A, 0% strain)
InGaAs(60A,0.35% comp.)
GaAs (60A, 0% strain)
InGaAs(60A,0.35% comp.)
GaAs
InGaAs (a=5.67A) on GaAs (a=5.65A). Only 800A total InGaAs can be grown without the onset of dislocations
InGaAsP (60A, 0.4% comp)
InGaAsP (60A, 0.4% tensile)
InGaAsP (60A, 0.4% comp)
InGaAsP (40A, 0.4% tensile)
InGaAsP (60A, 0.4% comp)
InP
Alternating compressive and tensile strained InGaAsP grown on InP– No critical thickness limitation

Fig. 3.6 Strain and strain compensation, illustrated with typical quantum well stacks

```
since the lattice constant of the film is greater than that
of the substrate.
The critical thickness is 5.7222/(2 * 0.0102), or 234 Å.
```

3.6 Direct and Indirect Bandgaps

This chapter is intended to cover, mostly qualitatively, the use of semiconductor materials in lasing systems and a description of fundamental limits and constraints. Properties such as band gap and lattice constant are determined by the composition of the material, and thin films (though they can confine electrons and holes to very high density and facility lasing) have certain additional constraints, based on the amount of strain the material can tolerate. The very basic question we will address before completing this chapter is why some semiconductors can be lasers (such as GaAs and InP, and associated compounds) while others cannot (like elemental Si or Ge).

To answer this qualitatively, let us return to the discussion on band gap in Sect. 3.2, and delve a little bit deeper into what the band structure of a solid really means.

In this section, we take GaAs as an example of a direct bandgap semiconductor. In fact it is an important laser substrate, particularly for 980 nm pump lasers and shorter wavelengths (based on the GaAs/InGaAs/AlGaAs) material system. The substrate for longer wavelength materials (around 1.3 to 1.6 μm) is InP, but everything discussed about GaAs applies to InP as well.

3.6.1 Dispersion Diagrams

The fundamental perspective is that the energy levels in a system are given by the solutions to Schrodinger's equation, Eq. 3.7.

$$\frac{-\hbar^2\nabla^2\psi}{2m} + U(x,y,z)\psi = E\psi. \tag{3.7}$$

An atom, for example, has discrete energy levels. These levels come out of Schrodinger's equation when the atomic potential (due to the protons at the nucleus) is put into the equation. (The energy levels which emerge predict all the atomic shells observed (*s*, *p*, *d*, *f,* and so on) and can be considered a major validation of quantum mechanics! These shells can be experimentally seen by exciting the atom with X-rays or electron beams, then watching the X-rays emitted from the excited atom.) In Fig. 3.7 is a schematic illustration showing how the energy levels in an atom become bands in a solid.

When this equation is applied to a three-dimensional periodic array of atomic potentials (a semiconductor crystal) the math gets complex, but the result is well known. The energy levels in the crystal become energy bands in the solid, with a band gap in between them. The significance of semiconductors is that each band holds four electrons/atom in the crystal, and semiconductors have a valence of four. This leads to a mostly empty band and a mostly fully band and all the desirable

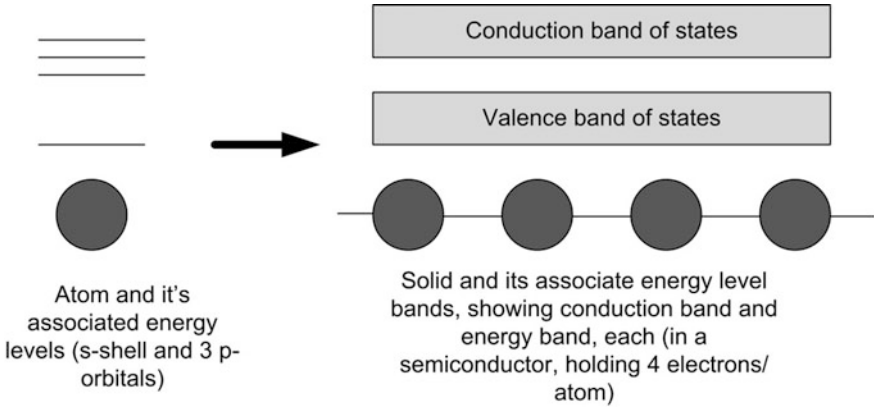

Conduction band of states
Valence band of states
Atom and it's associated energy levels (s-shell and 3 p-orbitals)
Solid and its associate energy level bands, showing conduction band and energy band, each (in a semiconductor, holding 4 electrons/atom)

Fig. 3.7 Atomic energy levels become energy bands when the atoms are placed in a three-dimensional crystal

properties of semiconductors, such as control of conductivity and carrier species (electrons or holes) through doping.

Schrodinger's equation has associated with each energy level E_n a k-vector (k_x, k_y, k_z). In 3D, solutions of the equation typically have a form $\exp(jk_x x + jk_y y + jk_z z)$, where k (as we discuss above) is fundamentally defined as $2\pi/\lambda$, where λ is the spatial wavelength in the direction specified.

An important dimension of the energy levels in a solid is how they depend on the k-vector. Intuitively, it makes sense that the electronic energy depends on the wavelength and direction associated with the electron in material. Electrons traveling in different directions interact with the crystal in different ways.

Usually, this relationship is captured in a dispersion diagram, which encapsulates the relationship between E and k in several different directions and will illustrate why Si and Ge are not good semiconductors.

Figure 3.8 illustrates a real space and reciprocal space, version of a unit cell of GaAs (which is a cubic lattice). The real-space version gives the dimension of the unit cell; the reciprocal space illustrates the appropriate k-vector associated with electronic wavelengths from 0 (delocalized) to $2\pi/a$ (localized to the crystal).

The special points labeled in Fig. 3.8 are the zone center (Γ, gamma point), face center X (chi), and corner (L) point. Typical dispersion diagrams for cubic semiconductor systems show E versus k starting with $k = 0$ and going toward both X and L. The *dispersion diagram* of a semiconductor captures the E versus k dependence of the solid. Since k includes direction, the dispersion diagram is plotted as a function of direction. The graph in Fig. 3.9 shows E versus k for GaAs, where the k-vector starts at 0 (a delocalized electron with a very long wavelength), and heads toward the center of the face of crystal (X) (indicate by Miller indices as the (100) direction, and toward the corner of the crystal (L), in the (111) direction). The key point of this diagram is the energy depends both on the magnitude of k and the direction associated with the carrier. The other major substrate for optoelectronics, InP, looks much like GaAs; it has heavy and light hole bands, a split-off band, and is a direct band gap.

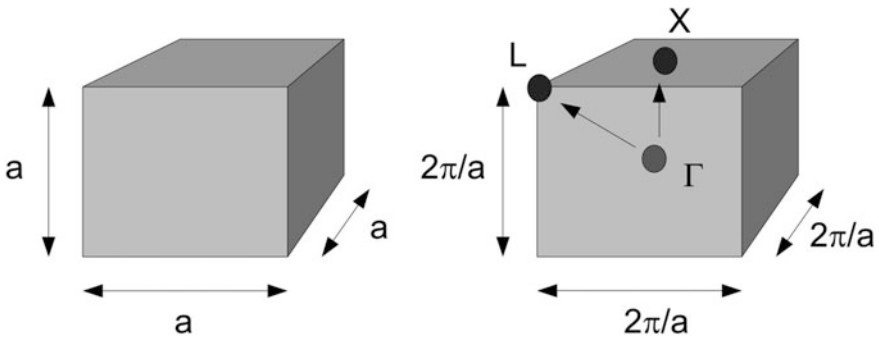


Fig. 3.8 Right, a real-space lattice picture showing a unit cube (shown in more detail in Fig. 3.3). Left, the reciprocal space picture, in which each dimension is drawn in units of $2\pi/a$. The dispersion diagram shown in Fig. 3.9 shows the E versus k curve, with k in the direction indicated

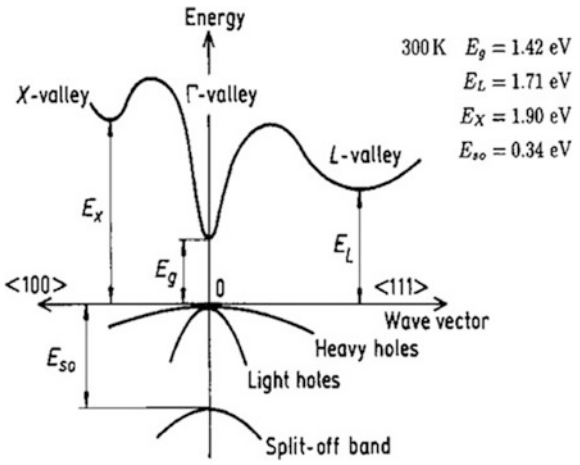


Fig. 3.9 Band structure of GaAs. Notice that there are several bands in the valence band, and that the band gap differs at different k values. From *Handbook Series on Semiconductor Parameters*, M. Levinshtein, S. Rumyantsev, M. Shur, ed., © 1996, World Scientific Pub. Co. Inc., used by permission

Note these are only typical directions in a crystal—there are many others, and some may be of interest, particularly considering transport in a given direction. However, they give a picture of the E versus k curve and illustrate the fundamental difference between direct bandgap semiconductor and indirect semiconductors. Usually, what we are most concerned with is the smallest distance between the highest valence energy level and the lowest electronic energy level—the band gap. Since electrons (and holes) settle to their lowest energy state, this is where most of the carriers will be and between where recombinations will take place.

3.6.2 Features of Dispersion Diagrams

The dispersion diagram has much more useful information than just the band gap. First, let us take a look at the conduction band of GaAs, shown in Fig. 3.9. The conduction band has various energies depending on direction and magnitude of k, but the lowest energy is at zone center ($k = 0$, or λ very large—a delocalized electron). Most electrons injected in a GaAs semiconductor will have a k value near 0, since that value corresponds to their lowest energy point.

The valence band has an interesting structure—in fact, it has three bands, known as the heavy hole band, the light hole band, and the split-off band. These bands all have slightly different density of states, associated effective masses of the carriers in the band, and even band gaps (as we will quantify in the next chapter). In practice, the material will be dominated by the lowest energy band with the highest density

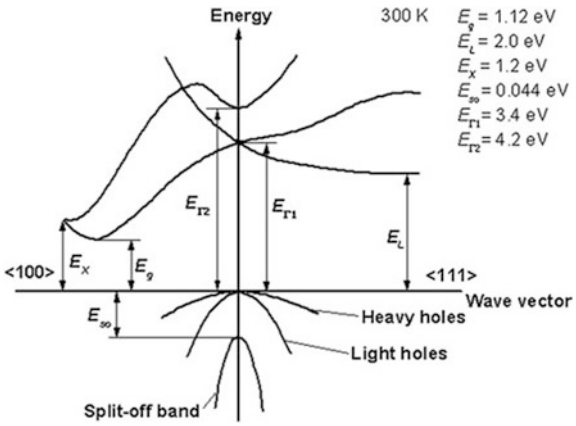


Fig. 3.10 Band structure of Si. The figures show that the minimum in the conduction band lies in *L* direction, toward a face. From *Handbook Series on Semiconductor Parameters*, M. Levinshtein, S. Rumyantsev, M. Shur, ed., © 1996, World Scientific Pub. Co. Inc., used by permission

of states (which, as we will see in the next chapter, is the heavy hole band in GaAs). Information about the density of states is actually in the E versus k curve as well.

This band structure is characteristic of unstrained GaAs. If a semiconductor is strained, some of the symmetries are broken, and the heavy and light hole bands are no longer at the same energy. Breaking this degeneracy between the heavy and light hold bands increases the differential gain and hence, speed of the device, as we will see in Chap. 8.

Many of the III–V semiconductors, particularly InP, have similar band structures.

3.6.3 Direct and Indirect Band Gaps

In the valence band, holes float up. Most of the holes will be also at zone center—the minimum in the conduction band is directly above the minimum (hole) energy in the valence band. This is crucially important for a laser material for the following reason.

Qualitatively, both electrons and holes have momentum associated with them, and momentum, like energy, needs to be somehow conserved in an interaction. The momentum associated with an electron or hole (or photon) in a crystal is given by the de Broglie relation

$$p = \hbar k. \tag{3.8}$$

When a recombination event occurs, an electron changes from a state in the conduction band to a state in the valence band, resulting in a net change of momentum, $\hbar\Delta k$, and a change in energy about equal to the band gap. The energy is

taken up by the emitted photon, but the emitted photon has very little momentum. In order for momentum to be conserved in a radiative recombination, either Δk has to be zero, or momentum has to be conserved some other way (through, e.g., lattice vibrations (phonons) which are discussed in Sect. 3.6.4). Involving three elements (an electron, hole, or phonon) makes this radiative recombination much less likely.

This requirement that Δk equal zero requires that the semiconductor be a direct-band-gap material, with the minimum in the conduction band being directly above the minimum (hole) energy in the valence band. In practice this means that $k = 0$ for both electrons and holes.

Semiconductors like GaAs and InP, and most of their heterostructures, such as InGaAsP, are direct band gap semiconductors, where valence band and conduction band energies have minima at the same k-value. The semiconductor Si, whose dispersion diagram is shown in Fig. 3.10, is not a direct band gap material. As can be seen, the conduction band minimum does not overlap the valence band (electron) minimum at $k = 0$. Therefore, Si can never be a good laser or light-emitting device, no matter how developed process technology or how inexpensive and available Si wafers become. Forever, we are doomed to expensive and beautiful III-V substrates.

Interestingly, Si can be an excellent light detector. When absorbing light, momentum is conserved by the interaction of phonons (lattice vibrations); as the light is absorbed, in addition to the generation of electron-hole pairs, lattice vibrations in the atoms are created (or absorbed). This process is much more efficient for absorption than for recombination, and so Si can detect light without being able to readily generate light.

3.6.4 Phonons

The lattice vibrations mentioned in the previous section are called phonons and they serve a useful role in allowing some recombinations and absorptions between carriers of different k-values. A semiconductor crystal consists of a bunch of atoms bonded together, but at temperatures above 0, each of these atoms is vibrating a bit about its equilibrium position. As the temperature increases, the atomic vibrations increase. These lattice vibrations serve to soak up excess momentum in many carrier-light interactions.

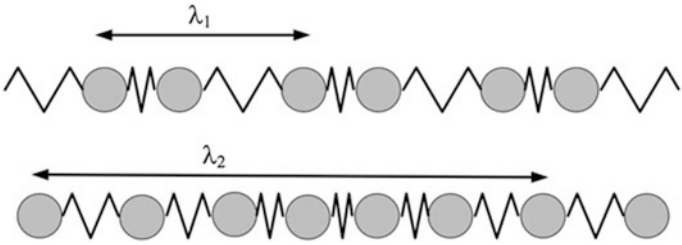


Fig. 3.11 Short wavelength and long wavelength phonons

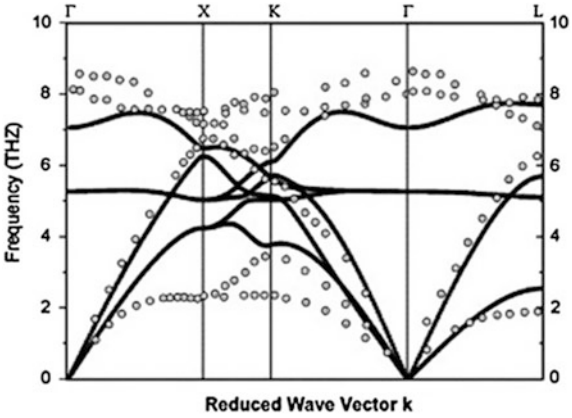


Fig. 3.12 Spectrum of phonons in GaAs, showing wide range of k's (x-axis) over very small energies (y-axis). Note the range of 10THz corresponds to an energy of 40 meV. From *Journal of Physics and Chemistry of Solids,* J. Cai, X. Hu, N. Chen, v. 66, p. 1256, 2006, used by permission

One useful conceptual picture is to imagine the atoms bonded atom-to-atom by little springs. As one atom vibrates, it pushes the atom next to it a bit away from its equilibrium position, which pushes on its neighbors, and so on. The picture is illustrated in Fig. 3.11. Now, the vibration becomes a crystal-wide phenomena with its own wavelength and k-vector, and the E versus k curves of these vibrations can be plotted.

The phonon band structure for GaAs is given in Fig. 3.12. Note the scale of the x-axis. These phonon vibrations have fairly low energy ($\sim$30 meV in GaAs), but span the entire range of k-vector, and hence, momentum.

An absorption event in Si is facilitated by phonon interaction. A 700 nm photon is absorbed in Si, transitioning with a Δk of about 2/3 π/a. The change in system momentum is taken up by either an optical phonon emission, resulting in an absorption energy $\sim$30 meV below or optical phonon absorption, resulting in an absorption energy $\sim$30 meV above 1.77 eV (the energy equivalent of 700 nm).

3.7 Summary and Learning Points

A. The wavelength of light which is emitted from a semiconductor wafer depends on the bandgap of the material and is given by 1.24 eV-μm/E_g(eV) = λ (μm).
B. The family of III-V semiconductors made with Ga, As, In, P, Al, and some other materials, can be made into heterostructures (like $In_{0.25}Ga_{0.75}As$), whose properties (like band gap, refractive index, and lattice constant) are (approximately) the weighted average of the binary constituents.
C. Because properties are roughly the weighted average of binaries, (the InP/InGaAsP) material system can access wavelengths spanning the

telecommunications range (from 1.3 to 1.6 μm) and still be lattice matched to an InP substrate.

D. Know Fig. 3.2 (the graph of the binary III-V compounds lattice constant and band gap)!
E. Lasers are made up of thin layers (quantum wells) stacked on one the other. Stacking material with mismatched lattice constants creates strain (distortion of the layer) or dislocations (missing atomic bonds).
F. Dislocations are fatal to lasers. It is very important that the layers be grown so as to minimize dislocations.
G. There is a critical thickness above which dislocations are created and below which the thin layer is strained.
H. Critical thickness limitations can be overcome by strain compensation.
I. Dispersion diagrams express the E versus k dependence of carriers and phonons in semiconductors. The propagation constant k is related to the momentum of the carrier or phonon, either electron or hole.
J. GaAs and InP are direct bandgap semiconductors, which readily emit light. Si and Ge are indirect bandgap semiconductors, which do not readily emit light and cannot in general be used for lasers.
K. A direct band gap semiconductor has its minimum electron energy exactly over the minimum hole energy on an E versus k diagram. Recombination between an electron and hole (emitting a photon) will involve no change in momentum. This is necessary, because photons carry very little momentum!
L. Phonons are quanta of lattice vibrations. In absorption of light in indirect band gap materials, they ensure that moment and energy are conserved.

3.8 Questions

Q3.1. What property of a semiconductor determines the wavelength of photons emitted by a particular semiconductor?
Q3.2. What is the name of the process by which semiconductors emit light?
Q3.3. Look at Fig. 3.2. What is the lattice constant (in Å) for InP? What is the wavelength corresponding to the energy gap for InP? What is the corresponding energy band gap in eV?
Q3.4. Look at Fig. 3.2. Suppose a semiconductor were made out of In, Al, Ga, and As. Estimate the range of energies the band gap could span and the range of lattice constants that it could span (hint: look at the properties of the binaries).
Q3.5. Why is an InP-based laser particularly useful for optical communications with optical fibers?
Q3.6. True or False. As the mole fraction of In increases in $In_{1-x}Ga_xAs$, the mole fraction of Ga decreases.
Q3.7. What is Vegard's Law? What is it used to calculate?
Q3.8. What is a thin film? How thick is a thin film (typically, in nm)?

Q3.9. What is the lattice constant of a material?
Q3.10. What is the strain of a material?
Q3.11. Define in your own words the critical thickness of a semiconductor.
Q3.12. True or False. A thin film grown on a material will be strained if its lattice constant is different than the substrate on which it is grown.
Q3.13. True or False. Dislocations can occur at the when thin films are grown on bulk material and serve to relieve strain at the interfaces.
Q3.14. True or False. As the lattice mismatch between a thin film and a substrate decreases, the strain exhibited in the thin film also decreases.
Q3.15. What is a typical value of a strain of a thin film in a semiconductor laser (%)?
Q3.16. True or False. As the degree of strain increases, the critical thickness decreases.
Q3.17. What is a direct bandgap semiconductor? List two examples.
Q3.18. What is an indirect bandgap semiconductor? List two examples.
Q3.19. True or False. As the value of the propagation constant increases (for an electron or hole or photon), the value of the momentum increases.
Q3.20. What is a phonon?
Q3.21. Explain in your own words how indirect band gap semiconductors, like Si, can absorb light while conserving energy and momentum.

3.9 Problems

P3.1. The refractive index of GaAs is 3.1, with a band gap of 1.42 eV. The refractive index of InAs is 3.5, with a band gap of 0.36 eV. (a) Find the composition of $In_xGa_{1-x}As$ that has a refractive index of 3.45. (b) Find the band gap at this composition.
P3.2. The data below gives data about the InGaAlAs system.

Compound	Bandgap (eV)	Lattice constant (Å)
InAs	0.36	6.05
GaAs	1.42	5.65
AlAs	2.16	5.66

An $In_{0.5}Ga_{0.3}Al_xAs$ quantum well is grown on InP (a = 5.89 Å).

(a) What is x?
(b) Estimate the band gap of the quantum well, treating it as a bulk material.
(c) What is the strain of this material when grown on InP (magnitude, and compressive or tensile)?

(d) Estimate how thick it can be grown without dislocations.

P3.3. Using the data of Table 3.1, find the composition of an $In_xGa_{1-x}As_yP_{1-y}$ alloy that has a band gap of 1.560 μm and a strain of +1% when grown on InP. (Note: while it is certainly possible to do this analytically, use of a spreadsheet or Matlab may facilitate a much quicker solution.)

P3.4. As noted in Sect. 3.6.4, phonons mediate absorption of light in indirect band gap materials. Because of this, materials can actually absorb from wavelengths 'slightly' below the band gap, due to phonon absorption. Qualitatively sketch the absorption coefficient of Si (E_g = 1.12 eV) keeping in mind that a absorption can take place slightly below the band gap, and that slightly above the band gap, two mechanism for photon absorption (involving phonon emission and phonon absorption) are available.

P3.5. (This problem is adapted from Kasap[1] and is used by permission). Figure 3.2 represents the band gap E_g and the lattice parameter a in the quaternary III-V alloy system. A line joining two points represents the changes in E_g and with composition in a ternary alloy composed of the compounds at the ends of that line.

The compound semiconductor $In_{0.53}Ga_{0.47}$ As has the same lattice constant as InP and can be alloyed with InP to obtain a quaternary alloy, $In_xGa_{1-x}As_yP_{1-y}$, whose properties lie on the line between $In_{0.53}Ga_{0.47}$ As and InP. Therefore, they all have the same lattice parameter as InP but different bandgaps.

(a) Show that quaternary alloys are lattice matched when $y = 2.15(1 - x)$.

(b) The band gap energy E_g in eV for $In_xGa_{1-x}As_yP_{1-y}$ lattice matched to InP is given by the empirical relation,

$$E_g(\text{eV}) = 1.35 - 0.72y + 0.12y^2$$

Calculate the fraction of As suitable for a 1.55 μm emitter.

[1] S. O. Kasap, *Optoelectronics and Photonics: Principles and Practices*. Upper Saddle River, NJ: Prentice Hall, 2001.

4 Semiconductors as Laser Materials 2: Density of States, Quantum Wells, and Gain

'If it cannot be expressed in figures, it is not science, it is opinion…"
—Robert A. Heinlein

Abstract

In the previous chapter, we discussed the direct properties of semiconductors that are relevant to lasers, including band gap, strain, and critical thickness. In this chapter, we talk about the ideal properties of semiconductors and semiconductor quantum wells, including density of states, population statistics, and optical gain, and we develop quantitative expressions for these that are based on ideal models. These will lead up to a qualitative and quantitative expression of optical gain.

4.1 Introduction

The general idea of semiconductor lasers formed by quantum wells which confine the carriers to facilitate recombination was described in Chap. 3, along with the various features of the band structure that facilitate recombination (direct vs. indirect band gap) and the limits on strained- and unstrained-layer growth of quantum well layers. However, to really focus on the precise effect of material, composition, and dimensionality (bulk vs. quantum wells vs. quantum dots) on optical gain, we need to develop expressions for carrier density and carrier properties. In this chapter, we develop a quantitative basis for carrier density, and optical gain in reduced dimension structures which will let us quantitatively understand the benefits of quantum wells (and other reduced-dimensionality structures) for lasing. By the end of the chapter, we will understand optical gain in terms of current density in a semiconductor.

© Springer Nature Switzerland AG 2020

D. J. Klotzkin, *Introduction to Semiconductor Lasers for Optical Communications*,
https://doi.org/10.1007/978-3-030-24501-6_4

4.2 Density of Electrons and Holes in a Semiconductor

In this chapter, we are going to drop, briefly, the reality of semiconductors and just consider an ideal semiconductor. We would like to determine the dependence of electron (and hole) density in a semiconductor as a function of energy and temperature. This energy band function is going to be critical in determining the optical gain of a semiconductor.

The first step is to calculate the density of electronic states. Here, the logic is identical to that we used in Chap. 2 in finding the density of states of photons in a black body. Take a cube of length L of semiconductor sitting in space, and consider which wavelengths λ or propagation vectors k will fit precisely in that cube.

The original assumption is that an electron, just like a photon, has an allowed wavelength that fits precisely into this imaginary cube of semiconductor material.

$$\begin{aligned} \lambda_{\text{allowed}-x,y,z} &= \frac{L}{m_{x,y,z}} \\ k_{\text{allowed}-x,y,z} &= \frac{m_{x,y,z} 2\pi}{L} \end{aligned} \tag{4.1}$$

The difference between this derivation and the photon derivation is the *changed energy -versus k* relation for electrons versus photons. For photons (as in Chap. 2), Planck's constant relates energy and optical frequency or wavelength, as in $E = h\nu = \hbar ck$.

For electrons, the relationship is different. One of the fundamental ideas of quantum mechanics is wave–particle duality: electrons are particles, having both mass m and an energy E; and waves, with a wavelength λ (or propagation constant $k = 2/\lambda$). In free space, the energy is related to the propagation constant k with the expression

$$E = \frac{\hbar^2 k^2}{2m} = \frac{1}{2} m v^2 = \frac{p^2}{2m} \tag{4.2}$$

This comes from de Broglie's relationship between wavelength and momentum of a particle with mass, mentioned in Chap. 3 and repeated here[1]:

$$p = \hbar k \tag{4.3}$$

The above equation is the fundamental description of a particle wavelength. From those two equations, we can obtain the k. versus E relationship for a particle (like an electron or hole) to be:

[1]This idea was put down in deBroglie's Ph.D. thesis. Would that you would have a thesis of such significance!

$$k = \frac{\sqrt{2mE}}{\hbar} \tag{4.4}$$

As in Chap. 2, the differential density of points in k-space is the volume in k-space

$$V(k)dk = 4\pi k^2 dk \tag{4.5}$$

divided by the volume of one point in k-space

$$V_{\text{allowed state}} = (2\pi/L)^3 \tag{4.6}$$

giving a number of points in k-space equal to

$$D(k)dk = \frac{4\pi k^2 dk}{(2\pi/L)^3} = \frac{L^3 k^2 dk}{2\pi^2}. \tag{4.7}$$

For each point in k-space, we need to multiply by a factor of two to represent the two electronic spin states (and hence, two electrons) for each state. To write Eq. 4.7 in terms of energy, we need expressions for both k and dk in terms of energy. Differentiating Eq. 4.4 we obtain

$$dk = \frac{2m\ dE}{\hbar\sqrt{2mE}} \tag{4.8}$$

Plugging in k and dk in terms of energy back into Eq. 4.7, and then dividing by L^3 (to get the density of states per unit real space volume), we obtain,

$$D(E)dE = \frac{(2m)^{\frac{3}{2}}E^{\frac{1}{2}}}{2\pi^2\hbar^3}dE \tag{4.9}$$

We have gone through this discussion rather quickly because we want to talk more about the physics rather than the math, and it closely echoes the density-of-state discussion of the photons in the black body.

The important thing is the physics that Eq. 4.9 expresses. In a three-dimensional, bulk crystal, the density of states is proportional to both the square root of the energy and the (effective) mass of the carrier, to the 3/2. Later, we will compare this to the density of states in a thin slab of material (a quantum well) and see one of the important advantages that these quantum wells possess.

4.2.1 Modifications to Eq. 4.9: Effective Mass

Equation 4.9 has mass in it. The E-versus-k (or E-vs.-λ) formula in a semiconductor crystal is more complicated than the free-space electron, because electrons or holes with varying effective wavelengths interact in different ways with the periodic atoms in the crystal (see Sect. 3.6). The potential energy term, involving the interaction of the charge carrier and the atomic cores, is very dependent on the k-value of the charge carrier.

Inside a crystal, the allowed energy is modified from the free space description above Eq. 4.2 by the presence of the atoms of the crystal. *However, the formula for density of states is essentially correct if we replace the free electron mass m with an effective mass m*.* This effective mass includes the effect of the crystal on the electrons in a single-lumped number. This approximation is especially true toward the bottom of the band gap where most of the carriers are. All the details of the interaction can be neglected with the net effect of being in a crystal replaced by a modification to a single mass number.

The effective mass is defined by the E-versus-k curve as

$$\frac{1}{m^*} = \frac{1}{\hbar^2}\frac{\partial^2 E}{\partial k^2} \tag{4.10}$$

This definition holds for any direction (x, y, and z) and any value of E. The dispersion diagram, effective mass, and density of states are all essentially descriptions of the same thing.

If the E-versus-k curve on the dispersion diagram is sharper, the effective mass of those carriers is lighter. Take a look, for example, at the dispersion diagram for GaAs in Fig. 3.9 in the previous chapter. The effective mass for electrons in GaAs is about 0.08 times the electron rest mass, and the effective mass for holes is about 0.5 m_0. This is clear from the dispersion diagram: at zone center ($k = 0$), the conduction band curvature is much sharper than the valence band, which is why conduction band electrons are much lighter. Consequently (because the density of states is proportional to mass), the density of states in the conduction band is much lower.

The effective mass defined in Eq. 4.10 depends on the direction of **k**, and there are effective masses for each direction. In addition, there are different effective masses appropriate for conduction (involving the application of outside fields) and for density of states/population statistics (in Eq. 4.9) which do not involve a particular direction. In the valence band, there are several bands (heavy hole and light hole) for the carriers to occupy, and each of these also has a different effective mass.

The effective mass for conduction in general is given by

$$\frac{3}{m^*_{\text{conduction}}} = \left(\frac{1}{m^*_l} + \frac{2}{m^*_t}\right), \tag{4.11}$$

where m_l and m_t are the E-versus-k masses in directions parallel, and transverse to, the appropriate minimum energy valley, respectively. For example, in Si, where the minimum energy is in the (100) direction, the longitudinal direction is (100), and the transverse directions are the (011) direction. This expression effectively averages the effective mass. In direct band gap semiconductors, with the minimum energy at $k = 0$ (a delocalized electron), the effective mass for conduction and density of states is simply the effective mass.

The effective mass for density of states (Eq. 4.9) does not involves a direction. It is given by the geometric mean of the effective masses in longitudinal and transverse directions as below.

$$m^*_{\text{density}-\text{of}-\text{states}} = (m_l m_t^2)^{\frac{1}{3}} \tag{4.12}$$

The situation is more complicated in the valence band, where there are several sub-bands, each of which can contain carriers (see discussion in Fig. 4.1). In Eq. 4.10, the term $\frac{\partial^2 E}{\partial k^2}$ is a function of the particular band $E(k)$ to which we are

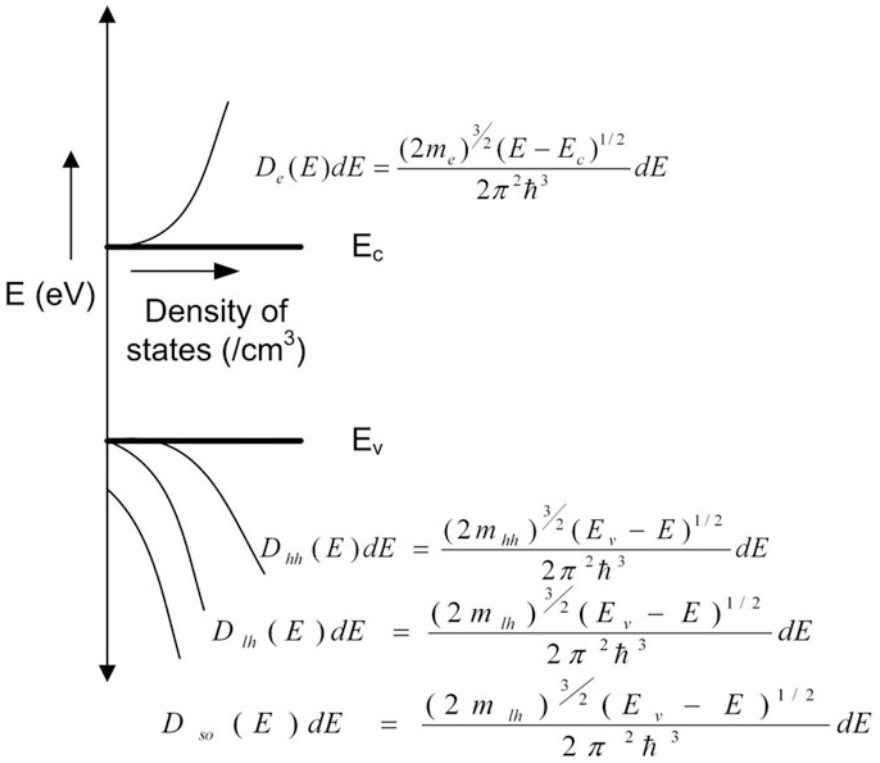


Fig. 4.1 Qualitative picture of density of states for both electrons and holes in GaAs, showing conduction band and valence light and heavy hole bands and the split-off band

referring. For example, the heavy hole effective mass depends on the curvature of the heavy hole band.

Combining the effective masses of the various bands in the valence band requires another average. Few carriers are in the split-off band because it is higher in energy than the other two bands. The appropriate average of the heavy hole and light hole bands for calculating the hole effective mass is

$$m_{\text{density-of-states}}* = (m_{hh}^{3/2} + m_{lh}^{3/2})^{\frac{2}{3}} \tag{4.13}$$

The central point here is that the effective masses used for equations for population statistics, and for conduction, are *appropriate averages* of the effective masses determined by the curvature of *E*-versus-*k* curves. For laser applications, the effective mass for conduction does not matter much, since the speed of the device is not determined by carrier transport. Instead, the effective mass for population statistics influences things like threshold current density and the like. However, in high-speed electronics, effective mass for conduction is the critical parameter, and it is for that reason that electronics designed for higher frequency operation (like GHz receivers for cell phones) typically uses Ge- or GaAs-based semiconductors which have much lower effective-mass carriers (particularly electrons).

One quick example will illustrate these calculations

Example: In Ge, with an energy minimum at 0.66 eV in the (111) direction, the electron transverse and longitudinal effective masses are
$m^*_{e,l} = 1.64$
$m^*_{e,t} = 0.082$
Estimate the effective mass appropriate for population statistics and for conduction.

Solution: The conduction effective mass, given by Eq. 4.11, is $\frac{3}{m^*_{\text{conduction}}} = \left(\frac{1}{1.64} + \frac{2}{0.082}\right)$, or $m^*_{\text{conduction}} = 0.12\ m_0$. The density-of-state mass is given by Eq. 4.12, with $m^*_{\text{density-of-states}} = (1.64 * 0.082 * 0.082)^{1/3} = 0.22\ m_0$.

The take-away message of this section is that there is no single electron mass, but instead it depends on direction, band, and context (conduction or density of states). The above expressions relate the effective masses defined by Eq. 4.10 to the effective masses that could be experimentally extracted though cyclotron resonance measurements or conductivity measurements. For lasers, the relevant effective mass is density-of-state effective mass.

4.2.2 Modifications to Eq. 4.9: Including the Band Gap

In addition, the density of states is zero in the band gap of the semiconductor crystal, and there are different density of states expressions for the electron and the holes. Shown below is a modified version of Eq. 4.9 in Fig. 4.1, along with a sketch of density of states, to correctly express this relationship.

Because the density of states is a function of mass, the density of states is lower for bands with lower effective mass. For example, in GaAs systems, the curvature of the conduction band is much sharper than the valence band, and therefore, the effective mass of electrons is lighter and the density of states is lower in the conduction band.

The valence band of GaAs is actually composed of three bands: the 'heavy hole,' 'light hole,' and 'split-off' bands (Figs. 3.9 and Fig. 4.1). The heavy hole band has a lower curvature, higher effective carrier masses, and larger density of states. Taking this one step further, because the heavy hole band does have much more room for carriers, most of the holes will be in the heavy hole band, and the properties of the holes in GaAs or other III–V materials tend to be dominated by the properties of this band.

The third band, the 'split-off' band is at slightly higher energy than the other two and does not generally contain many free carriers.

All of the details and complexity of the band structure come about from the detailed solution of Schrodinger's equation for a very complex atomic potential. That particular problem is beyond the scope of the book, but in Sect. 4.3, we look at the solution of the very simple potential represented by a quantum well structure.

This is a good description to the density of states in a bulk semiconductor. We need to apply this to the quantum well structures that are commonly used in semiconductor lasers.

4.3 Quantum Wells as Laser Materials

Let us introduce a quantum well and demonstrate its importance to semiconductor lasers.

A quantum well is a thin slice of material of a lower band gap, sandwiched between two other materials of larger band gap. These energy walls confine the carriers (electron and holes) to stay mostly in the well. In fact, this real 'particle in a well' is an excellent analogy to the classical quantum mechanical example of a particle in a well.

Figure 4.2 shows both a schematic picture of a well, with the electrons and holes confined to the slab, a sketch of an electron microscope picture of a laser, showing materials with different composition forming a set of multiquantum wells (which is how most lasers are formed) and an scanning electron microscope image of a set of quantum wells. The particle in a box is out of its box! It is now a useful engineering construct.

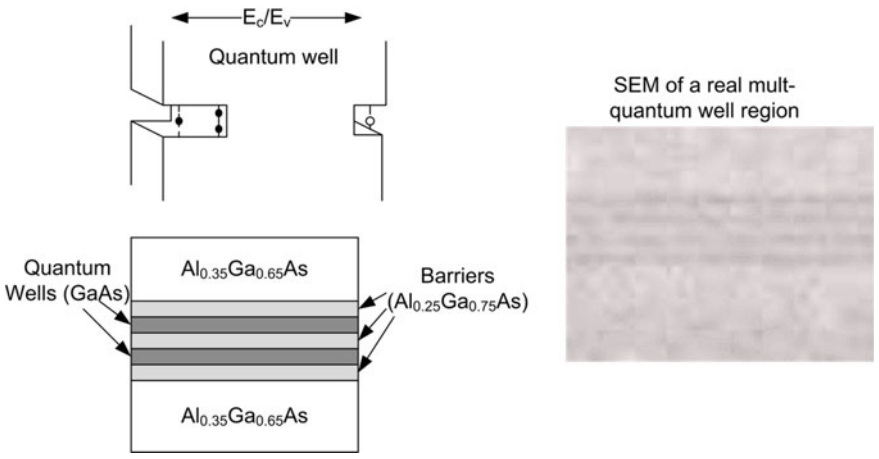


Fig. 4.2 Above left, a picture of a single quantum well, showing how the electrons and holes are confined in the quantum well giving rise to quantized energy levels. Below left, a schematic of a multiquantum laser, showing individual wells, separated by barriers. Right, a scanning electron microscope image showing quantum wells in an actual laser. Almost all semiconductor lasers are multiquantum well lasers

These semiconductor quantum wells form confining potentials (or 'little boxes') in which carriers (electrons and holes) are trapped. Because they are confined by the energy barriers around them, the density of electrons and holes in the same location is much higher than it would be otherwise. This enhancement of carrier density is critical in realizing high-performance semiconductor lasers.

It is really impossible to overstate the importance of quantum wells in modern semiconductor lasers. It is quite difficult to make a working laser at high temperature with a bulk semiconductor material. The unconfined carriers and light would require much higher current densities to lase. Compared to a bulk *p–n* junction with the same current input, the carrier density in the quantum well is much higher and all of the performance characteristics are much better.

Let us now quantify a bit more what happens to the density of states, and energy levels, in a quantum well.

4.3.1 Energy Levels in an Ideal Quantum Well

Let us first look at the energy levels in an ideal quantum well of width *a,* and solve directly for the energies and wavefunctions of that system, pictured below in Fig. 4.3.

In Chap. 3, Eq. 3.7 expressed Schrodinger's equation in a three-dimensional form. Here, we would like to solve the one-dimensional form of Schrodinger's equation, where Ψ is the wavefunction, U is the potential energy function, and E_n is the energy eigenvalues.

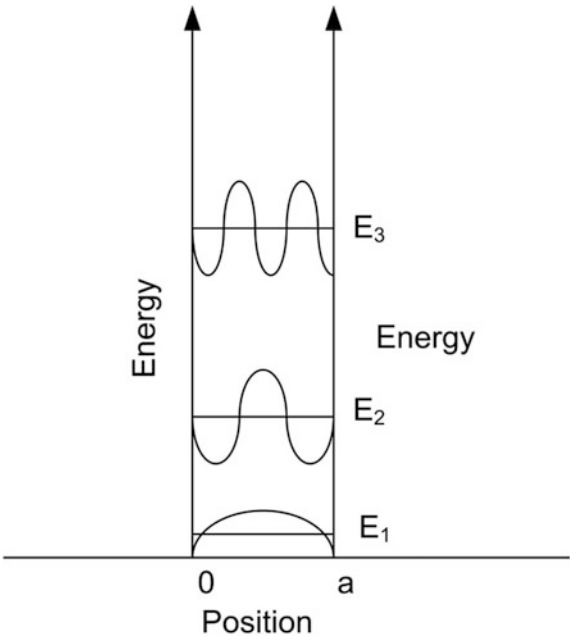


Fig. 4.3 Picture of the energy levels and wavefunction of a particle in an infinite quantum well. Outside of the region from 0 to a, the energy barriers are infinite, and the particle is constrained to remain in that range from 0 to a. The lines show the energy levels, and the curves indicate the wavefunctions associated with them

$$\frac{-\hbar^2\nabla^2\psi}{2m} + U(x)\psi = E_n\psi \tag{4.14}$$

This equation can be used to give a very good model to what a quantum well does to the energy band structure of a semiconductor.

The potential profile of the ideal quantum well above has its potential energy as $U = 0$ between $x = 0$ and $x = a$, and infinite (with the particle forbidden) outside that range. The wavefunction Ψ is required to be continuous at the boundary 0 and a, and the appropriate boundary conditions are that the wavefunction and its derivative equal 0 at the boundaries of the well.

For this simple case, Schrodinger's equation can be written as

$$\frac{\hbar^2\nabla^2\psi}{2m} = E\psi \tag{4.15}$$

inside the well, which has a solution of the form

$$\Psi(x) = A\sin(k_z z) \tag{4.16}$$

where A is a currently undetermined constant. This expression is always zero at $x = 0$, and equals 0 at $x = a$ if $k_z a$ is an integral multiple of π, or

$$k_z a = n\pi \tag{4.17}$$

Equation 4.17 defines k_n and the only remaining variable is A. To evaluate a value for A, recall that the interpretation of the wavefunction is that $\Psi^*\Psi$ yields the probability density at a particular location in the spatial domain. Thus, the integral of $\Psi^*\Psi$ over the entire permissible domain should be equal to 1, requiring that particle should be somewhere. Mathematically,

$$\begin{aligned} 1 &= \int_0^a A^2 \sin^2(n\pi z) dz = \frac{aA^2}{2} \\ A &= \sqrt{\frac{2}{a}} \end{aligned} \tag{4.18}$$

(To simplify evaluating the integral, we recall that the average of $\sin^2(x)$ or $\cos^2(x)$ over any number of integral half periods is equal to ½, and so evaluating the integral is just multiplying this average by the width of the range (a in this case). This sort of integral is ubiquitous, so it is worthwhile to remember and apply!).

We now know exactly what the wavefunction $\Psi(x)$ is. By substituting this into Eq. 4.15, above, we can obtain the allowed energy values (or energy eigenvalues) that are allowed by Schrodinger's equation. We get energy eigenvalues of

$$E_n = \frac{n^2\hbar^2\pi^2}{2ma^2} \tag{4.19}$$

Because the particle is confined, the energies of the confined particles are lifted above the ground state bulk level by E_n. The narrower the well is, the greater the lift. This one-dimensional confining potential acts like an artificial atom with discrete energy levels. The steps in the energy level are proportional to the quantum number, n, squared.

4.3.2 Energy Levels in a Real Quantum Well

Let's take a two-step approach to understanding a real semiconductor quantum well, illustrated in Fig. 4.4. First, what happens when the confining potential is non-infinite and exists for both electrons and holes? Qualitatively, the result is essentially the same. Energy levels appear in the quantum wells. As these energy levels rise higher and higher, they eventually escape the confining potential and then appear as part of the bulk density of states in the 'barrier' region around the quantum well. Because the mass of electrons and holes is different, the energy levels and offsets are different in the valence and conduction bands. In addition, for a real quantum well (say, a semiconductor quantum well with a band gap of 1 eV in a 'barrier' region with a cladding of 1.2 eV, as pictured), the total confining potential of 0.2 eV divides up in different way between the valence and conduction band depending on the material system. (This topic will be discussed in a later chapter).

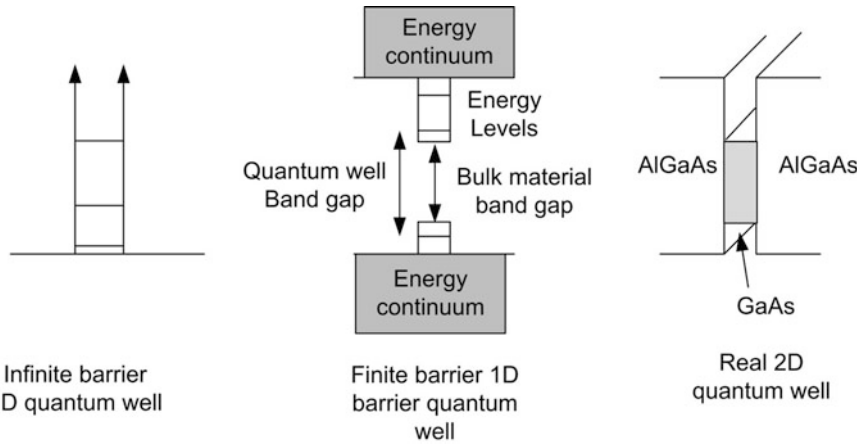


Fig. 4.4 Left, an ideal quantum well in 1D with infinite barriers; middle, a finite 1D quantum well with barriers for both the electrons and holes; and right, a real semiconductor quantum well, showing finite barriers, an unconstrained k_x and k_y and a k_z constrained by the quantum well. In these figures position is shown on the 'x'-axis, and energy is shown on the 'y'-axis

Because recombination happens between electron and hole states, effectively, in a quantum well, the band gap is higher than that in the bulk material. The effective band gap is between the first hole level and the first electron level, as seen in Fig. 4.4b.

Let us do a real example to calculate the magnitude of this effect.

Example: A layer of InGaAsP with a bulk material band gap of 1.3 μm is confined in a quantum well of 80 Å width. The effective mass of holes is 0.6 m_0 and of electrons is 0.08 m_0. Estimate the emission wavelength of this quantum well.

Solution: The energy level corresponding to 1.3 μm is 0.954 eV. From Eq. 4.19, the approximate shift in the valence band is

$$\Delta E = \frac{1^2(1.05 \times 10^{-34})^2(3.14)^2}{2(0.6)(9.1 \times 10^{-31})(80 \times 10^{-10})^2} = 1.55 \times 10^{-21}\, J = 0.010\,\text{eV}$$

and similarly, in the conduction band is $\Delta E = 0.072\,\text{eV}$. As shown in the picture below, these offsets add to the bulk band gap to produce a net band gap of 0.954 + 0.010 + 0.072 = 1.034 eV, and a corresponding recombination wavelength of 1.20 μm.

The effect is illustrated pictorially in Fig. 4.4. The quantum wells formed in both the valence and conduction bands shift the band gap up and lower the emission wavelength from the bulk value.

4.4 Density of States in a Quantum Well

In the beginning of Sect. 4.3, we described qualitatively why quantum wells aid enormously in laser performance. To quantify this statement, we need to develop the expression for density of states in a quantum well.

Shown in Fig. 4.5 is a picture of a very thin slab of material (a quantum well) as well as a picture of its density of states, in k_x and k_y, in k-space. Let us first calculate the density of states, in states/cm^2 (*not* cm^3) in this thin slab of material. This is strictly a calculation in two dimensions.

Then, we can include the k_z values permitted by Eq. 4.17 to generate a sketch of states/cm^3, including the thickness of the material.

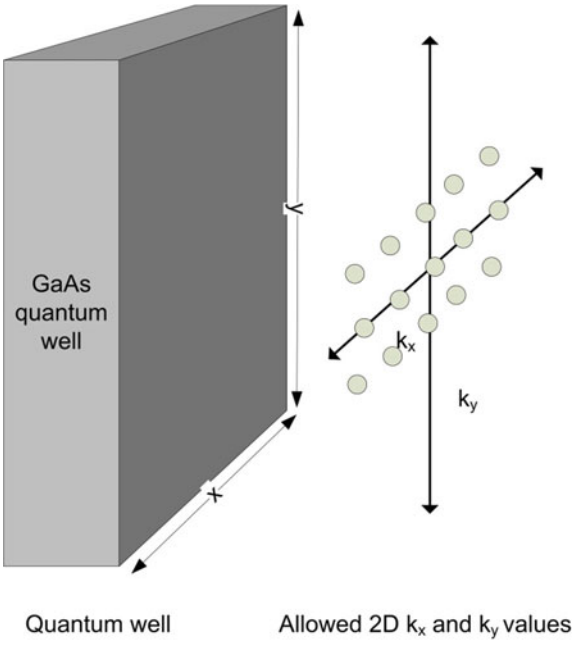


Fig. 4.5 A schematic picture of a quantum well, showing a thin z and large (macroscopic) x and y. Next to it a 2D k-space picture, showing allowed k-values in k_x and k_y

As before, the boundary condition is assumed to be that the wavefunction equals 0 at $y = x = L$, in a 2D square of dimension L. The areal density of states A_d picture now is the fraction of points inside a circle of radius k, or the *area* in k-space

$$A_d(k)dk = 2\pi k dk \tag{4.20}$$

Divided by the area of one point in k-space

$$A_{\text{allowed state}} = (2\pi/L)^2 \tag{4.21}$$

or a areal density of points in k-space, we obtain

$$A_d(k)dk = \frac{2\pi k\ dk}{(2\pi/L)^2} = \frac{L^2 k dk}{2\pi} \tag{4.22}$$

There are two spin states allowed for each electronic state. Using the expressions for energy versus k and dk in Eqs. 4.4 and 4.8, and multiplying by two to account for the two spin states, the areal density of states for a quantum well as a function of energy per cm^2 is

$$A_d(E)dE = \frac{m^*_{\text{density-of-states}}\ dE}{\hbar^2 \pi} \tag{4.23}$$

The interesting result is that the density of states is independent of energy. A careful look at the calculation will show that a 2D structure has the dimensionality so that the quadratic dependence of energy on propagation vector k just cancels the dependence of the density of k-points with increase of magnitude of k. The mass m^*_{dos} is the density-of-state effective mass.

This calculation, however, just captures the 2D density considering k_x and k_y. The sketch below expresses what happens when we include k_z and E_z in the third dimension. (These are given by Eqs. 4.17 and 4.19, respectively.) Since each k_z implies a fixed value of energy, the bottom of the band is offset by E_1. When the energy reaches the density associated with E_2, there are two values of k_z with the same density of states in k_z and k_y, and the net density of states doubles. These ideas are captured in the sketch of density of states sketch in Fig. 4.6, which compares a bulk semiconductor with a quantum well.

The importance of this abrupt step-like density of states, compared the gradual increase in density associated with the bulk, is that it causes a much higher carrier density at the band edge. For the same number of carriers injected, the carrier density at one particular energy will end up higher compared to a bulk semiconductor. Since the optical gain will depend in the carrier density at a given energy, having higher densities of carriers at one energy is clearly beneficial!

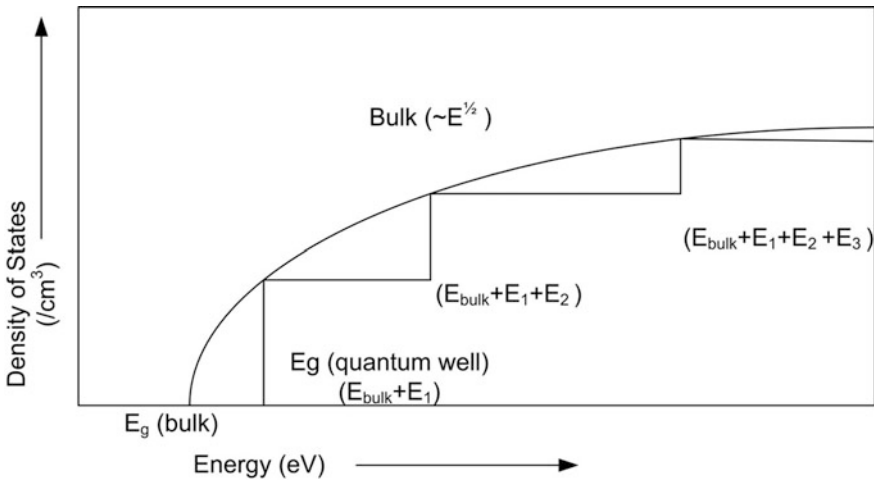


Fig. 4.6 A sketch of density of states of a quantum well vs. density of states for a bulk semiconductor material. The steps indicate sub-bands of the quantum well and are different values of k_z

4.5 Number of Carriers

The next thing we are interested in is the number of carriers (electrons or holes) in a given band. The basic expression in a bulk semiconductor is

$$n(E)dE = D(E)f(E, E_f)dE \tag{4.24}$$

where $n(E)$ is the number of carriers as a function of energy E, $D(E)$ is the density of states at an energy E, and $f(E, E_f)$ is the Fermi–Dirac distribution function as a function of the energy and the Fermi level E_f. We remind the reader that this Fermi function gives the probability that an existing state is occupied. From Chap. 2, the Fermi function is given as

$$f(E, E_f) = \frac{1}{1 + \exp((E - E_f)/kT)} \tag{4.25}$$

The idea of a 'Fermi level' E_f, is not fundamentally appropriate to lasers. Fermi levels are used to describe systems in thermal equilibrium, and as we discussed in Chap. 2, lasers cannot be in thermal equilibrium. They have to be driven by some non-equilibrium means (typically electrical injection for semiconductor lasers) in order to be put into a state of population inversion. However, the expressions above are still used with the introduction of *quasi-Fermi levels*.

4.5.1 Quasi-Fermi Levels

Equation 4.25 above still has some utility with regard to lasers. Even though the electrons and holes are not in thermal equilibrium with each other, we can assume that the electron population and hole population are separately in thermal equilibrium, but each with a different 'quasi-Fermi level'. The concept is illustrated in Fig. 4.7.

The figures on the left show semiconductors in true thermal equilibrium in an *n*-doped semiconductor. If the Fermi level is near the top (say, by *n*-doping), there are lots of electrons in the valence band and a few in the conduction band. If the Fermi level is near the bottom in a *p*-doped quantum well, there are lots of holes but very few electrons. The second figure from the left in Fig. 4.7 shows a true thermal equilibrium in a *p*-doped system.

The third figure from the left represents a *p*–*n* junction with a forward bias applied which is *not* in thermal equilibrium. A separate 'quasi-Fermi level' for electrons, E_{qfe}, and holes, E_{qfh}, describes the population density in the conduction and valence band, respectively. When we are calculating the density of electrons in the conduction band, the quasi-Fermi level for electrons is used, and when calculating the density of holes, the quasi-Fermi level for holes is used.

The figure in the far right represents a p–n junction under strong forward bias, where the quasi-Fermi levels for electrons and holes are no longer in the band gap, but are actually in the bands. This situation has a very high density of electrons and holes in conduction and valence band, and is actually what is necessary for lasing.

The distribution of the electrons in the conduction band is still assumed to be 'equilibrium'. They interact with each other, and their distribution among the available density of states is thermal and determined by the Fermi distribution function. However, the number of electrons is determined by the quasi-Fermi level. The mental picture is that a large number of electrons are electrically injected into

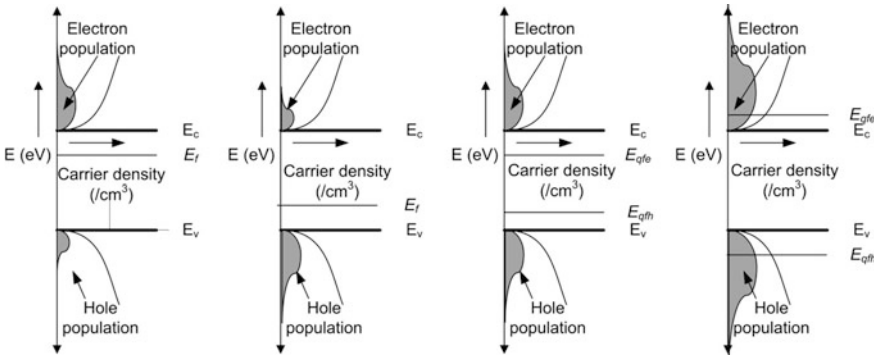


Fig. 4.7 Illustration of the distribution of carriers as a function of Fermi level (left) and two separate 'quasi-Fermi levels' right. The situation on the far right has a high number of both electrons and holes

the conduction band of the quantum wells from the *n*-side of the junction, where they then interact with each other, and with the lattice of atoms, and quickly distribute themselves thermally. Similarly, holes are injected from the *p*-side of the junction, and then distribute themselves thermally as well. In this picture, the quasi-Fermi level is a shorthand description of the number of carriers in the band.

4.5.2 Number of Holes Versus Number of Electrons

To avoid potential confusion, let us write down the separate expressions for density of holes and density of electrons. The Fermi–Dirac expression gives the probability of an *electron* state being occupied. The probability of it being vacant, or occupied by a hole, is $1–f(E, E_f) = f(-E, -E_f)$. The density of states for holes increases as energy decreases (hole energy increases as electron energy decreases). Typically, we are interested in hole populations below the Fermi level of interest where $E–E_f$ is negative. The combination of all these expressions gives the expression for density of holes as a function of energy. The appropriate functions for holes and electrons are given in Table 4.1.

A good way of visualizing it is that for holes, the energy should be read as increasing downward—that is place a negative sign in front of every energy value, and, since only differences between energies appear, calculations will work out correctly.

4.6 Condition for Lasing

At this point, we have expressions for the density of electrons and the numbers of their respective quasi-Fermi levels. What electron and hole density do we need for lasing?

Table 4.1 Electron and hole density for bulk semiconductors

	Electrons	Holes
Appropriate Quasi-Fermi level	E_{qfe}	E_{qfh}
Distribution Function	$f_e(E, E_{\mathrm{qfe}}) = \frac{1}{1+\exp((E-E_{\mathrm{qfe}})/kT)}$	$f_h(E, E_{\mathrm{qfh}}) = \frac{1}{1+\exp((E_{\mathrm{qfh}}-E)/kT)}$
Density of states	$D_e(E)dE = \frac{(2m_e)^{\frac{3}{2}}(E-E_c)^{\frac{1}{2}}}{2\pi^2\hbar^3}dE$	$D_h(E)dE = \frac{(2m_h)^{\frac{3}{2}}(E_v-E)^{\frac{1}{2}}}{2\pi^2\hbar^3}dE$
Number of carriers	$n_e(E)dE = \frac{1}{1+\exp(\frac{E-E_{qfe}}{kT})}\frac{(2m_e)^{\frac{3}{2}}(E-E_c)^{\frac{1}{2}}}{2\pi^2\hbar^3}dE$	$n_h(E)dE = \frac{1}{1+\exp(\frac{E_{qfh}-E}{kT})}\frac{(2m_h)^{\frac{3}{2}}(E_v-E)^{\frac{1}{2}}}{2\pi^2\hbar^3}dE$

As we talk about in Chap. 2, to achieve lasing, stimulated emission needs to dominate absorption:

$$BN_2N_p(E) > BN_1N_p(E) \xrightarrow{\text{implies}} \text{nonequilibrium system} N_1 < N_2 \tag{4.26}$$

where N_2 is the density of excited atoms, N_1 is the density of ground-state atoms, and $N_p(E)$ is the photon density at a particular energy E. There we were talking about discrete atoms states, where an atom by itself was either excited or in the ground state. We need to write this condition in terms of the population in the electron and valence band.

First, as mentioned in Sect. 3.6.3, photons carry very little change in momentum. For these optical transitions, Δk has to be 0. For any one particular electron energy E_{ec}, there is one matching valence band energy E_{ev} that has the same k, and the recombination between those two has a specific recombination energy E.

A reasonable assumption with which to start is that absorption is proportional to the number of electrons in the valence band, and the number of empty states (holes) in the conduction band. Since these are independent and independently given by the quasi-Fermi levels, the total absorption rate is proportional to the product of the two. Similarly, we assume that stimulated emission is proportional to the number of electrons in the conduction band and the number of empty states (holes) in the valence bands

$$\begin{aligned} \text{stimulated emission} &\propto f(E_{\text{ec}}, E_{\text{qfe}})(1 - f(E_{\text{ev}}, E_{\text{qfh}}))D_e(E_{\text{ec}})D_h(E_{\text{ev}}) \\ \text{absorption} &\propto f(E_{\text{ev}}, E_{\text{qfh}})(1 - f(E_{\text{ec}}, E_{\text{qfe}}))D_e(E_{\text{ec}})D_h(E_{\text{ev}}) \end{aligned} \tag{4.27}$$

in which E_{qfe} and E_{qfh} are the electron and hole quasi-Fermi levels, and E_{ev} and E_{ec} are the electron energy associated with a particular photon energy in the conduction and valence band, respectively.

For stimulated emission to be greater than absorption, with the expression above implies that

$$\begin{aligned} f(E_{\text{ec}}, E_{\text{qfh}})(1 - f(E_{\text{ev}}, E_{\text{qfe}}))D_e(E_{\text{ev}})D_h(E_{\text{ec}}) &> f(E_{\text{ec}}, E_{\text{qfe}})(1 - f(E_{\text{ev}}, E_{\text{qfh}}))D_e(E_{\text{ec}})D_h(E_{\text{eh}}) \\ f(E_{\text{ev}}, E_{\text{qfh}}) &> f(E_{\text{ec}}, E_{\text{qfe}}) \end{aligned} \tag{4.28}$$

With a little algebra, the expression above can be rearranged to show

$$E_{\text{ec}} - E_{\text{ev}} < E_{\text{qfe}} - E_{\text{qfh}} \tag{4.29}$$

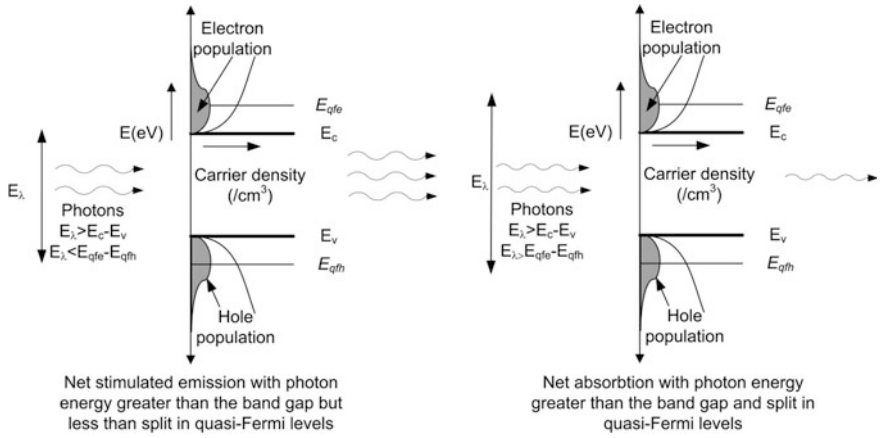


Fig. 4.8 Bernard–Duraffourg condition. At the left, photons incident on a semiconductor with an energy greater than the band gap but less than the split in the quasi-Fermi levels induce net stimulated emission, and possibly lasing. At right, higher energy photons are above the band gap, but experience net absorption, rather than stimulated emission

In order for stimulated emission to be greater than absorption, and for lasing to be possible, the split in quasi-Fermi levels has to be greater than the laser energy levels! This condition is called the Bernard–Duraffourg condition after the people who first described it in 1961. It is illustrated in Fig. 4.8.

Not only are semiconductor lasers not in equilibrium, but they are very far from equilibrium. The split between quasi-Fermi levels (which, we recall, is zero in equilibrium) must be at least as great as the band gap (the minimum distance between electron and hole energies) in order for lasing to be possible in a semiconductor.

4.7 Optical Gain

It is only a short step from Eqs. 4.27 and 4.28 to an expression for optical gain. Let us first define optical gain as a measurable parameter and then write down the expression for optical gain in a semiconductor, including the ideas of density of states and quasi-Fermi levels that we have developed.

The term optical gain in a material means that when light is shined on it or through it, more light comes out than went in. Absorption of light is much more commonplace (everywhere from window shades to sunglasses) but optical gain has its important place in physics and technology. The erbium-doped fiber amplifier,

which allows optical transmission over thousands of miles, is based on optical gain and can amplify signals by factors of 1000.

Phenomenologically, optical gain and absorption are described by the following equation.

$$P = P_0 \exp(gl) \tag{4.30}$$

where P_0 is the initial optical power, P is the final power, and the 'gain' g is in units of length^{-1} and is positive for actual gain and negative for absorption. (Typically, in laser contexts, gain and absorption are expressed in units of cm^{-1}). Two quick examples will suffice to illustrate this formula.

Example: About 95% of the power is transmitted through window glass 1 cm thick. What is the absorption coefficient of window glass, and what fraction of a 100 W light beam will make it through the window?

Solution: $P/P_0 = 0.95 = \exp(g1)$, so g = ln(0.95) = −5.1 cm^{-1}, or an absorption of 5.1 cm^{-1}.

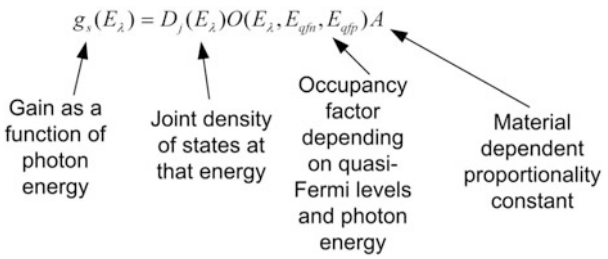
Example: An erbium-doped fiber amplifier has a gain of about 30 dB over a length of about 3 m of fiber. What is the gain in cm^{-1}? If the input is 1 μW, what is the output power?

Solution: A gain of 30 dB means 30 = 10 log (P/P_0), so $P/P_0 = 1000 = \exp(-g3000)$ and g = ln(1000)/3000 = 0.0023 cm^{-1}. The output power gain of 30 dB means that the output increases by a factor of 1000, giving an output power of 1 mW.

4.8 Semiconductor Optical Gain

Finally, let us write down an expression for the optical gain in a semiconductor, as a function of material properties, density of states, and quasi-Fermi levels. This expression will capture the dependence of gain on carrier injection level, degree of quantum confinement, and material properties.

The simple optical gain expression consists of the product of three separate terms, representing three different factors. They are the density of possible recombinations (which is known as the 'joint,' or 'reduced' density of states, discussed below); occupancy factor related to the charge density, determined by the quasi-Fermi levels for electrons and holes; and a proportionality factor (amount of gain for each possible absorption or recombination state). These terms are written in Eq. 4.31.


$g_s(E_\lambda) = D_j(E_\lambda)O(E_\lambda, E_{qfn}, E_{qfp})A$
Gain as a function of photon energy
Joint density of states at that energy
Occupancy factor depending on quasi-Fermi levels and photon energy
Material dependent proportionality constant

(4.31)

Finally, there is a linewidth broadening factor, which includes small variations from strict k-conservation and allows recombination between electrons–holes of slightly different k-values.

4.8.1 Joint Density of States

Let us look at the graph in Fig. 4.9, showing the process of recombination under conditions of strict k-conservation. The energy of the emitted photon is given by the band gap plus the offset in both the valence and conduction bands. With strict k-conservation, any particular photon energy E_λ has exactly one k-value associated with that recombination.

The E-versus-k relationship for photon energy is then given by the expression below.

$$E_\lambda = E_g + \frac{\hbar^2 k^2}{2m_e} + \frac{\hbar^2 k^2}{2m_h} = E_g + \frac{\hbar^2 k^2}{2m_r} \tag{4.32}$$

with the term, m_r, defined as thereduced mass,

$$\frac{1}{m_r} = \frac{1}{m_e} + \frac{1}{m_h} \tag{4.33}$$

These two equations lead to a photon energy E_λ-versus-k relationship for the photons of

$$k = \frac{(2m_r(E_\lambda - E_g))^{1/2}}{\hbar} \tag{4.34}$$

Just as in considering density of states for electrons and holes, every allowed k-value constitutes a state. Here, each single value of k represents a single allowed transition. Hence, the density of possible photon emissions (called reduced density

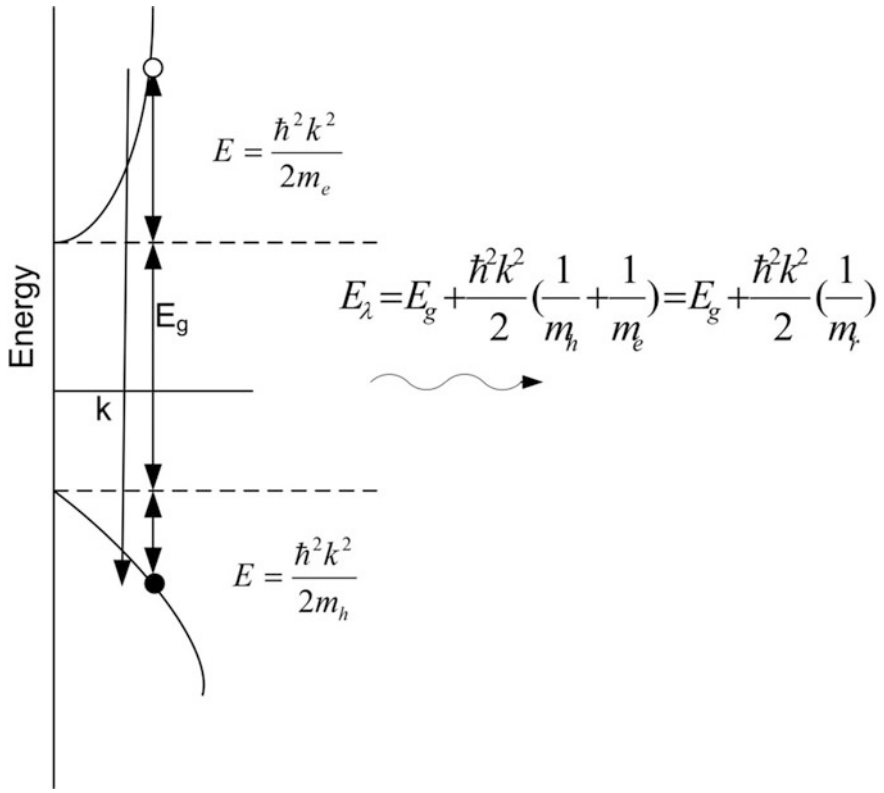

Energy
Eg
k
$E = \frac{\hbar^2 k^2}{2m_e}$
$E_\lambda = E_g + \frac{\hbar^2 k^2}{2}\left(\frac{1}{m_h} + \frac{1}{m_e}\right) = E_g + \frac{\hbar^2 k^2}{2}\left(\frac{1}{m_r}\right)$
$E = \frac{\hbar^2 k^2}{2m_h}$

Fig. 4.9 Relationship between photon energy E_λ, band gap energy E_g, and k. The large down arrow illustrates the recombination, which emits the photon, while the two smaller arrows indicate the distance from the band edge

of states or joint density of states) is given by the same process used for density of electron states, with the slightly modified E-versus-k relationship given in Eq. 4.35,

$$D_j(E_\lambda)dE = \frac{(2m_r)^{3/2}(E_\lambda - E_g))^{1/2}}{2\pi^2\hbar^3}dE \tag{4.35}$$

This joint density of states term is one part of the gain expression, and represents the density of transitions for a given photon energy E_λ.

4.8.2 Occupancy Factor

Of course, just as an electronic state either has an electron it or not, the joint density of states has to be appropriately populated in order to provide gain or absorption. Let us think about a 'recombination state' of fixed photon energy E_λ. There exist a

number of electrons, which can participate in this recombination (all of those at the corresponding electron energy). The fraction of possible electrons which can participate is given by the Fermi function, $f(E_{qfe}, E_{ec})$, and the fraction of possible holes is given by the number of vacant electronic states in the valence band, $1-f(E_{qfv}, E_{ev})$. The total number of 'gain states,' proportional to each is proportional to the product, $f(E_{qfe}, E_{ec})\ (1-f(E_{qfv}, E_{ev}))$. (As in Sect. 4.5, E_{qfx} is the appropriate hole or electron quasi-Fermi level, and E_{ec} and E_{ev} are the energy levels which satisfy k-conservation for a given recombination energy and wavelength E_λ.)

Similarly, the total number of absorption states is proportional to the product of the number of vacant electronic sites at the appropriate conduction band energy level and the number of occupied electronics states in the appropriate valence band energy level, $f(E_{qfe}, E_{ec})\ (1-f(E_{qfv}, E_{ev}))$.

The net occupancy factor is proportional to this total number of gain states minus the number of absorption states, or

$$\begin{aligned} O &= f(E_{qfc}, E_{ec})(1 - f(E_{qfv}, E_{ev}) - f(E_{qfv}, E_{ev})(1 - f(E_{qfc}, E_{ec}) \\ &= f(E_{qfc}, E_{ec}) - f(E_{qfv}, E_{ev}) \end{aligned} \tag{4.36}$$

This argument is illustrated pictorially in the simple band diagram of Fig. 4.10. The figure shows a single conduction and valence band level, both of them appropriate for recombination for a particular photon energy E_λ. The net gain is related to the difference between the number of recombination states indicated by down arrows and absorption states indicated by up arrows. In the figure shown, the relevant electron level has $f(E_{qfe}, E_{ec}) = 0.66$ and the relevant hole level has $f(E_{qfv}, E_{ev}) = 0.33$.

First, if both states contain a hole, or both an electron, then no recombinations are possible. To get gain, we need population inversion, which means an electron in the conduction band and a hole in the valence band.

4.8.3 Proportionality Constant

The most effective way to write down this 'proportionality constant' between the number of available transitions and the gain in cm^{-1}, is to start with the final answer. The expression for gain can be written down as

$$g(E_\lambda)dE = \frac{(2m_r)^{\frac{3}{2}}(E_\lambda - E_g)^{\frac{1}{2}}}{2\pi^2\hbar^3} \times (f(E_{qfc}, E_{ec}) - f(E_{qfv}, E_{ev})) \times \frac{\pi\hbar q^2}{6\varepsilon_0 cm_0 n_r E_\lambda} f_{cv} \tag{4.37}$$

It is a monstrous beast of an expression, but the origin of the first two parts should be clear, and the last part is the proportionality constant A. In the expression, ε_0 is the free space dielectric constant and n_r is the relative permittivity of the semiconductor. The term f_{cv} is related to the *quantum mechanical oscillator*

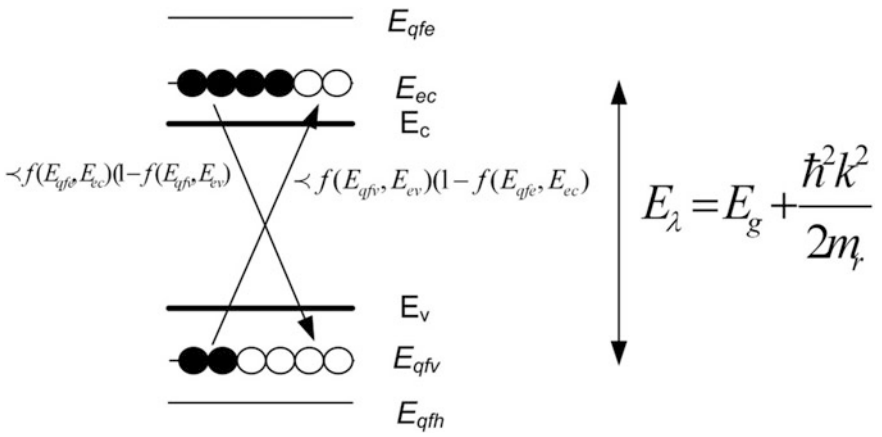


Fig. 4.10 Illustrating the occupancy factor *O*, which is the difference between the relative number of recombinations and absorptions. Only one conduction and valence band level participate in a radiative recombination at a particular photon energy level

strength of the transition of the electron from the conduction to the valence band, which represents how likely a recombination is to take place. It can be taken as a material constant, with a value of 23 eV in GaAs for an allowed ($\Delta k = 0$) transition and 0 for a forbidden ($\Delta k <> 0$) transition.

If properly evaluated with consistent units, the equation gives gain in units of length^{-1}.

Recall also that E_{qfc} and E_{qfv} are alternative ways of expressing the carrier density, and E_{ev} and E_{ec} are not independent energy values, but are uniquely specified by the photon energy E_λ.

4.8.4 Linewidth Broadening

Looking at the gain formula in Eq. 4.37, we can see that is largely composed of the density-of-state term for the system we are observing. Hence, for a bulk system, we expect to see something that varies quadratically with energy and for a quantum well system, we would expect to see an abrupt increase in gain right at the first quantum well energy level transition and another abrupt increase in gain when the energy hits the second allowed transition (depending on carrier populations).

That is not, however, what is observed. The measured gain (which can be seen with a variety of techniques) is a very smoothed and softened version of what Eq. 4.37 predicts. The gain is convolved with a smoothing function, called a lineshape or a linewidth broadening function. This function serves to turn the theoretical sharp edges into smoother gradual rises. The effect of linewidth broadening on the gain spectrum can be seen in Fig. 4.11.

The physical origin of this function comes largely from violation of absolutely strict *k*-conservation due to scattering of the electrons and holes by phonons.

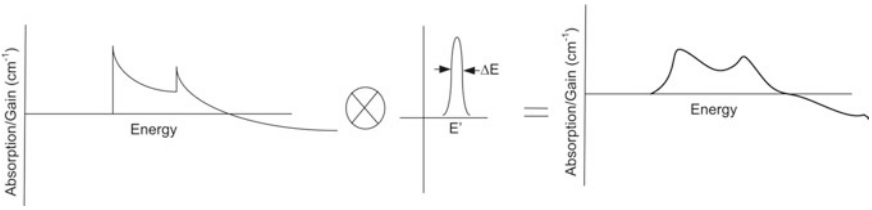


Fig. 4.11 A sketch illustrating the original gain expression, the lineshape function with which it is convolved, and the final (measured gain). The circled *X* represents the convolution operation. ΔE is characteristic of the width of the lineshape function, and the shape differs slightly depending on whether it is a Gaussian or Lorentzian

Should they interact, the energy conversation equation when the electron and hole recombine will include the energy of the phonon. Therefore, a single electron–hole recombination can emit a photon with a narrow range of energies, not just the exact wavelength set by the difference between hole and electron energy levels. If this interaction is uniform with all recombinations across the gain band, it is called homogenous broadening. If the phenomenon is specific to one range of wavelengths or one spatial area, it is called inhomogeneous broadening.

The new gain equation for this broadened gain $g_b(E_\lambda)$ then is given by the convolution of the lineshape function with the original function $g(E_\lambda)$

$$g_b(E_\lambda) = \int g(E_\lambda)L(E' - E_\lambda)dE' \tag{4.38}$$

where $L(E)$ is the appropriate lineshape function. The function is picked with a phenomenological linewidth and is normalized, so its integral is 1.

Two common forms are used for this lineshape function. The most common is called the Lorentzian lineshape function,

$$L(E' - E_\lambda) = \frac{1}{\pi}\frac{(\Delta E/2)}{(E' - E_{0\lambda})^2 + (\Delta E/2)^2} \tag{4.39}$$

where ΔE is the width of the linewidth function (often about 3 meV for these sort of models). This Lorentzian function is often used to model homogenous broadening.

Also used to model linewidth broadening is a Gaussian expression, such as

$$L(E' - E) = \frac{1}{\sqrt{2\pi}\Delta E}\exp^{-\frac{(E'-E_\lambda)^2}{2\Delta E^2}} \tag{4.40}$$

Finally, in this whole section, an expression for gain is developed as a function of material parameters and injection density. An interesting way to measure optical gain directly from analysis of the optical spectrum is presented in Chap. 7.

4.9 Summary and Learning Points

This chapter covers most of the common models and ideas that are used for semiconductor lasers, including benefits of quantum confinement, gain expression, quasi-Fermi levels, and Bernard–Duraffourg condition. With this foundation, it is hoped that most of the properties and experimental characteristics of lasers you encounter can be understood, modeled, and optimized.

4.10 Learning Points

A. The Pauli exclusion principle states that no two electrons can occupy the same quantum mechanical state or have the same quantum numbers.
B. The formula for density of states in a semiconductor gives the number of states available for electrons or holes at a given energy.
C. The density of states in a bulk semiconductor increases quadratically with energy, as $E^{1/2}$.
D. The two-dimensional density of states in a quantum well is constant. The sub-bands associated with the third dimension result in a staircase-like density of states versus energy.
E. The abrupt increase in density of states in a quantum well is very beneficial for lasing because it results in a lot of carriers at the same energy. Because of this, threshold current densities are much lower and semiconductor lasers are now almost universally quantum wells or smaller dimensions.
F. The number of carriers in a band at a given energy is given by the product of the density of states and the Fermi function.
G. Under conditions of electrical injection (or optical injection), the semiconductor is not under thermal equilibrium. In that case, the population of electrons and holes can be described by separate quasi-Fermi levels.
H. The quasi-Fermi levels are shorthand descriptions for the number and distribution of carriers in each band.
I. The lasing energy E_λ has to be less than the split between the quasi-Fermi levels in order for stimulated emission to dominate absorption.
J. Optical gain depends on the density of states (dependent on the dimensionality of the system and effective mass); the occupancy of holes and electrons (dependent on the quasi-Fermi levels; a proportionality constant; and a linewidth broadening factor.
K. This linewidth broadening factor is usually modeled as a Lorentzian or Gaussian expression with a phenomenologically determined linewidth.

4.11 Questions

Q4.1. What is the expression for the carrier density in a semiconductor? Explain what each of the terms (symbols) represents.
Q4.2. How does the density of states depend on the energy in a three-dimensional, bulk crystal, and in a 2D quantum well?
Q4.3 What is effective mass? Why is effective mass for density of states and conduction different?
Q4.4. What happens to the value of the effective mass as the curvature of the *E*-versus-*k* curve increases?
Q4.5. What is a quantum well? What is a quantum well composed of? Explain both the mathematics and the physical structure.
Q4.6. True or False. As the width of a quantum well increases, its energy levels decrease.
Q4.7. Will the energy offsets from the bulk band edge be greater in the conduction band or the valence band?
Q4.8. Will the luminescence wavelength of bulk $In_{0.3}Ga_{0.7}As$ or $In_{0.3}Ga_{0.7}As$ in a quantum well be longer?
Q4.9. What is the Bernard–Duraffourg condition?
Q4.10. What is optical gain?
Q4.11. What factors determine optical gain in a semiconductor?
Q4.12. Why are sharp gain edges, such as would be predicted by Eq. 4.37, not observed in gain measurements?

4.12 Problems

P4.1. Derive the density of states for a 1D quantum wire, in which the electrons are quantum-confined in two dimensions and free to move in only one dimension. The answer should be in units of length^{-1} energy^{-1}.
P4.2. A simple 3D model for the *E*-vs.-*k* curve around $k = 0$ is.$E(k) = A \cos(k_x a) \cos(k_y b) \cos(k_z c)$. What is the effective mass for density of states at $k = 0$?
P4.3. A 3D quantum box can be described as having a wave function of the form $\Psi(x, y, z) = A \sin(k_x x) \sin(k_y y) \sin(k_z z)$. If the box is a square box of dimension a,

(a) Write an expression for the energy level in terms of the quantum numbers, n_x, n_y, and n_z.
(b) Sketch the density of states for this system for the first four energy levels.

P4.4. In a certain semiconductor system, the density of states for electrons at $T = 0\ K$ is given in Fig. P4.12.

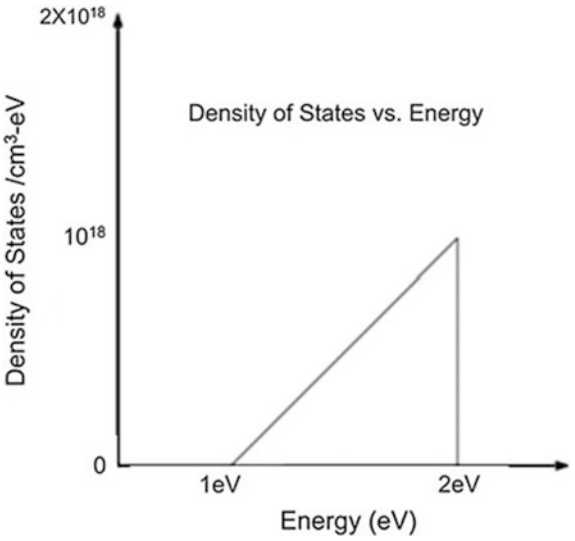


Fig. P4.12 Density of states of an odd semiconductor system

(a) If the system contains 2×10^{17} electrons/cm^3, what is the Fermi level?
(b) If the Fermi level is 0.8 eV, how many electrons does the system contain?
(c) Sketch the electron density versus energy at 300 K if the Fermi level is at 1.5 eV.

P4.5. Optical fiber has a loss of 0.2 dB/km. Calculate the loss in/km, and the power exiting the fiber after 100 km if the input power is 2 mW. (These are typical numbers for semiconductor optical transmission.)
P4.6. Calculate and graph the optical gain vs. energy for a simplified model of GaAs in which $m_e = 0.08m_0$, $m_h = 0.5m_0$, and $E_{qfv} = E_{qfc} = 0.1$ eV into their respective band, and $\Delta E = 3$ meV with a Gaussian lineshape function.
P4.7. Fig. 3.12 shows the band structure of Si.

(a) Sketch qualitatively the effective mass versus k of the lowest energy conduction band, indicating where it is negative, positive or infinite, from the < 000 > direction toward the < 100 > direction
(b) The valence bands include the heavy hole band, the light hole band, and the split-off band. Explain (briefly) which of these bands is most significant in determining the density of carriers versus temperature and Fermi level in the valence band.
(c) Estimate the longest wavelength that a Si photodiode can detect.
(d) Explain (briefly) how Si can absorb photons even though it is an indirect bandgap semiconductor.

P4.8. It is desired to make a 60 Å quantum well of InGaAsP with an emission wavelength of 1310 nm. If the effective mass of electrons is $0.08m_0$ and the effective mass of holes is $0.6\ m_0$, estimate the target emission wavelength of InGaAsP (considered as bulk semiconductor) to be grown, taking into account quantum well effects.

P4.9 A quantum dot is a small chunk of 3D material which has discrete energy levels. A quantum dot laser is made up of a collection of many, many of these dots, distributed in the active region. A simple model of a quantum dot has a single electron level and a hole level for each dot. A quantum dot active region has a number of dots in it, and the density of states given is given by the number of dots.

One of the implications of Eq. 2.15 is that the absorption coefficient is proportional to

$$\alpha = \alpha_0(N_2 - N_1)$$

where N_2 is the fraction of atoms in the excited state and N_1 is the number of dots in the ground state. Initially there is not current in the dots (N_1 = 1 and N_2 = 0). In this problem, light exactly matching the gap between the two levels is shined on an active region as pictured in Fig. P4.13.

(a) A very low level of light I_0 is shined on a quantum dot active region 1 mm long. The output light is 5×10^{-4} time the input light. Find α_0.
(b) A moderate level of light is shined on the active region, to maintain N_1 = 0.75 and N_2 = 0.5. If a small additional increment of light ΔL_{in} is shined on the active region, what is the increment of light out ΔL_{out}?
(c) If an enormous amount of light is shined on the active region ($L \rightarrow \infty$), what will N_1 and N_2 be? Is it possible to optically pump this region into inversion?

P4.10 Quantum dots, like atoms, have more than one electronic energy level. Suppose 100 quantum dots make up the active region of a quantum dot laser, as shown in Fig. P4.14. The first energy level is 0.1 eV above some reference, and the second energy level is 0.3 eV above the same reference.

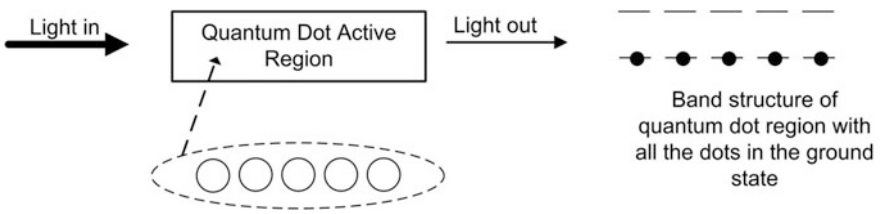


Fig. P4.13 A model of a quantum dot active region, showing left a range of dots inside of a structure, and right, the band structure of each dot, showing all the dots in the ground state

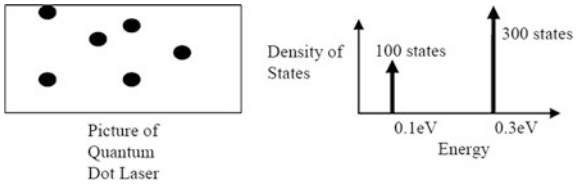


Fig. P4.14 Left, picture of arrangement of quantum dots inside the laser active region, and right, picture of density of states of quantum dots

Recall the Fermi occupation probability from Table 2.1 of Chap. 2.

(a) If the Fermi level is 0.05 eV below the first energy level at room temperature, how many of those energy states are occupied?
(b) If half of the energy states of the first energy level are occupied, what is the electron quasi-Fermi level?
(c) Why are there 300 states at the second energy level but only 100 states at the lowest energy level?
(d) What is the minimum number of electrons needed to get lasing from the first energy level (assuming that the number of holes injected into the valence band, not shown, is equal to the number of electrons in the conduction band)?

Check for updates

Semiconductor Laser Operation

5

… Rail on in utter ignorance
Of what each other mean,
And prate about an Elephant
Not one of them has seen!
—John Godfrey Saxe

Abstract

In the previous chapter, we talked about the ideal properties of semiconductors and semiconductor quantum wells, including density of states, population statistics, and optical gain, and develop expressions for these that are based on ideal models. In this chapter, we will take a step back to see how optical gain and current injection interact with the cavity and photon density to realize lasing. Finally, we present a simple rate equation model and examine it to see how laser properties such as threshold and slope are predicted. The predictions from the rate equation model are related to the measurements which can be made on these devices to determine fundamental properties of laser material and structure, including internal quantum efficiency and transparency current.

5.1 Introduction

In Saxe's famous poem, *The Blind Men and the Elephant*, six blind men discuss whether an elephant is like a rope, a fan, a tree, a spear, a wall, or a snake. The message at the end of the poem is that while each of them focuses on some aspect of the animal, they all miss the essentials of the elephant. Like an elephant, a semiconductor laser is several things. It is simultaneously a P-I-N diode (an electrical device) and an optical cavity, and both of these parts have to work together in order to be a successful monochromatic light source.

© Springer Nature Switzerland AG 2020

D. J. Klotzkin, *Introduction to Semiconductor Lasers for Optical Communications*,
https://doi.org/10.1007/978-3-030-24501-6_5

Rather than leaping into the study of the various parts of the laser, and ending up, like the men of Indostan in the poem, familiar with the parts but not the whole, in this chapter, we introduce a canonical semiconductor laser structure and describe it to the point where details about the waveguide, and the electrical operation and metal contacts can be sensibly studied in subsequent chapters. Let us look at the elephant before we dissect the poor thing!

5.2 A Simple Semiconductor Laser

Let us look again at the structure in Fig. 1.5. The single semiconductor bar serves as both a gain medium, as current is injected, and as a cavity, which confines the light.

In the latter part of Chap. 4, we discussed optical gain, and we saw that material with optical gain amplifies incident light. We also saw how a direct band gap semiconductor can exhibit optical gain if the hole and electron levels are high enough so that the quasi-Fermi levels are in their respective bands. All of this leads to a simple description of an optical amplifier, but it does not quite produce the clean, single-wavelength output of the ideal lasing system.

In Chap. 1, we saw that lasing requires a high photon density and gave examples of a HeNe laser in which the high photon density was achieved with mirrors which kept most of the photons inside the cavities. In the most basic semiconductor edge-emitting devices, the 'mirror' that keeps the photon density high inside the semiconductor optical cavity is formed by the cleaving of the semiconductor wafer. Since the dielectric constant of the semiconductor is typically around 3.5, and that of air, 1, the amplitude reflectivity r at the interface is given by

$$r = \frac{n_{\text{air}} - n_{\text{semi}}}{n_{\text{air}} + n_{\text{semi}}} \tag{5.1}$$

and the power reflectivity R (which is Eq. 5.1, squared) is

$$R = \left(\frac{n_{\text{air}} - n_{\text{semi}}}{n_{\text{air}} + n_{\text{semi}}}\right)^2 \tag{5.2}$$

For typical semiconductor laser values, R is about 0.3. These cleaved laser bars come with built-in mirrors that reflect 30% of the incident back into the cavity. This reflectivity is sufficient to achieve lasing in these structures. In general, the facets of commercial devices are also coated after fabrication with dielectric layers to increase (or reduce) their reflectivity at specific wavelengths.

5.3 A Qualitative Laser Model

Figure 5.1 is a picture of a qualitative laser model. It shows a collection of electrons and holes, which are electrically injected as current into the cavity. Let us imagine inside this cavity an optical wave bouncing back and forth between the mirrors, increasing exponentially according to the gain of the cavity, as it did at the end of Chap. 4. As the wave moves through the cavity, its intensity grows, due to the optical gain from the semiconductor. Let us ask the rhetorical question: Can the amplitude continue to grow without limit as it bounces back and forth?

The answer it is that it cannot: There is a feedback between the gain and the photon density that is important when the photon density is large. *Every photon which is created involves the removal of an electron and a hole*. As the photon density increases, the hole and electron density decrease, and the gain decreases. The laser is not just an optical amplifier, but an optical amplifier with feedback!

With this idea that an increase in photons leads to a decrease in gain (which leads in turn back to a decrease in photons) let us show why, under steady-state conditions, there has to be 'unity round trip gain' in a laser. Figure 5.1 shows optical modes inside of a laser cavity, growing exponentially as they travel back and forth, with many of the photons exiting from each facet. To anticipate later discussion and a potential difference in reflectivity between the two facets, the reflectivities at the two facets are labeled R_1 and R_2.

The term 'steady state' means that nothing changes with time; the injected current, and the carrier density and photon density inside the cavity look the same now as they did fifteen minutes ago or will fifteen minutes hence. The term 'unity round trip gain' in a laser means that the optical wave power after bouncing back

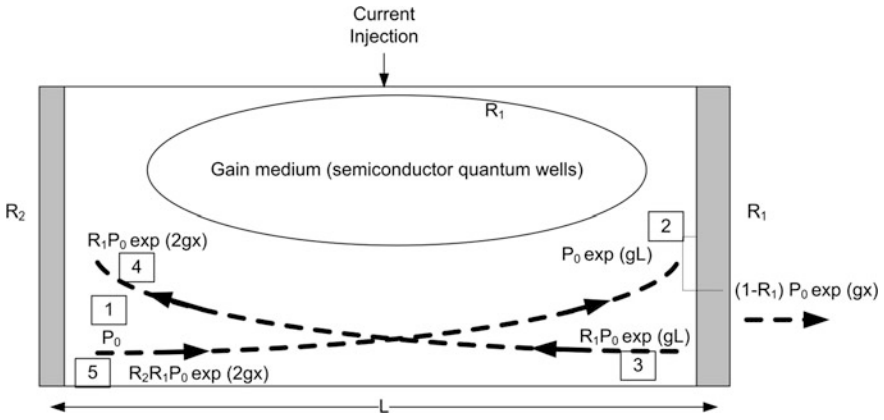


Fig. 5.1 Qualitative model of a semiconductor laser, showing optical waves propagating forward and backward, while gain is provided by carriers inside the cavity. Because of the feedback between the photons and the gain medium, there is required to be unity round trip gain, where $P_0 = P_0\ R_1R_2\exp(2gL)$

and forth between the cavity should be at the same level as when the wave started; the net gain, including power that leaks out of the facets, should be one.

In Fig. 5.1, we follow the path of the optical mode as it goes back and forth within the laser cavity. First, at position 1, the wave starts out with a value P_0 and increases exponentially according to the cavity gain g as it travels to the right facet. When it arrives there (position 2), on the right, its amplitude is $P_0\exp(gL)$. At the right facet, R_1 power is reflected, so the amplitude returning to the left is $R_1P_0\exp(gL)$. Finally, as the wave travels back toward the left, it experiences another cycle of exponential gain ($R_1P_0\exp(gL)\exp(gL)$, or $R_1P_0\exp(2gL)$) and another reflection $R_1R_2P_0\exp(2gL)$. That value, $R_1R_2P_0\exp(2gL)$, has to be equal to the initial photon density P_0, which sets the value of the gain.

Let's imagine what would happen if the gain were higher in Fig. 5.2. Then, after making one round trip, the optical wave would be a little larger. As the wave went around again, it would grow larger yet. Eventually, as the photon density in the cavity grows too large, the increased density would deplete the electrons and holes and reduce the gain. (The sharp-eyed reader may have already noticed that even under this condition, photons are constantly being created to replace the ones leaking out the facets; that is, the constant current coming in, which we are ignoring for the next two paragraphs, is just sufficient to replace the photons which are exiting the facets.) A similar argument can be made if the gain is lower than equilibrium; the carrier density would then build up to achieve unity round trip gain.

The value of the gain has to be such as to maintain the laser in steady state because of the interconnection between photon density and carrier density. The particular value of the equilibrium gain g depends on the cavity properties such as facet reflectivity. In Fig. 5.2, we show the photon density driving the gain, but of course it can be looked at the other way too; the gain drives the photon density. Regardless, the gain is a constant and is fixed in a lasing cavity to achieve unity round trip gain.

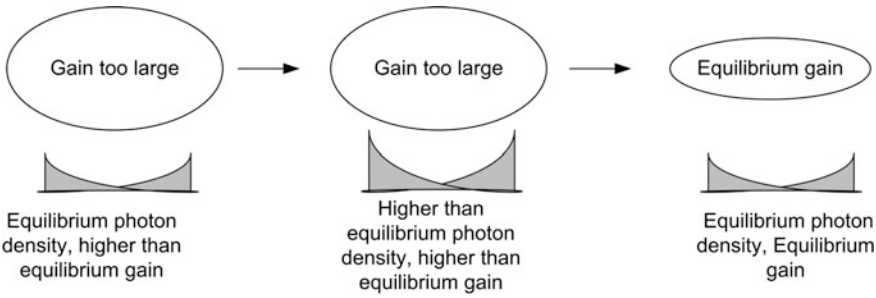

Gain too large
Gain too large
Equilibrium gain
Equilibrium photon density, higher than equilibrium gain
Higher than equilibrium photon density, higher than equilibrium gain
Equilibrium photon density, Equilibrium gain

Fig. 5.2 Feedback between the photon density and the gain. The oval represents the density of carriers which provide the gain. Read from right to left, this illustrates how if the gain is too large, it will eventually deplete carriers and reduce the gain back to its equilibrium value

The equation for unity round trip gain leads to the following relationship between cavity gain g_{cav} and facet reflectivity.

$$\begin{aligned} 1 &= R_1 R_2 \exp(2 g_{cav} L) \\ g_{cav} &= \frac{1}{2L} \ln\left(\frac{1}{R_1 R_2}\right) \end{aligned} \tag{5.3}$$

The steady state, DC, lasing gain is set by the condition of the cavity (*facet reflectivity and length*). Instead of analyzing the very detailed dependence of gain on quasi-Fermi level and band structure, we can simply look at the cavity length and reflectivity to determine an expression for the lasing gain.

For those with a background in electronics, the situation is analogous to the open and closed-loop gain of an op-amp or transistor. The 'open-loop' gain we studied in Chap. 4 was a function of the details of the band structure and semiconductor material system. The closed-loop gain of Eq. 5.3 depends on the feedback elements placed around it (in this case, the laser cavity). Like electronics, it is the closed-loop gain which is more important in setting device properties, though the intrinsic material gain sets limits.

The simplest useful model of semiconductor laser peak gain as a function of carrier density, or current density J, is given by the expression

$$g = A(n - n_{tr}) = A'(J - J_{tr}) \tag{5.4}$$

where n_{tr} is called the *transparency carrier density*, and J_{tr} is the *transparency current density* (both figure-of-merit material constants), and A and A' are proportionality constants with appropriate units. Let us define the carrier density at which a particular device starts to lase as n_{th}, the threshold current density. If we equate this to the cavity gain of (Eq. 5.3)

$$A(n_{th} - n_{tr}) = \frac{1}{2L} \ln\left(\frac{1}{R_1 R_2}\right), \tag{5.5}$$

it immediately says that the *carrier density is clamped to be* n_{th} *in a device which is lasing*. Because nothing on the right side of the equation depends on the current density, the value of the gain in the cavity cannot change with current density; therefore, the carrier population n is clamped at threshold to some population n_{th}. This expression is a more mathematical way of restating the discussion around Fig. 5.2. The photon density inside of the cavity (and exiting the laser) will vary, but the carrier density inside a laser cavity is fixed above threshold and is independent of the photon density. This idea will be revisited when we talk about the rate equation model for lasers and about their electrical characteristics.

Example: A bathtub has a hole in it. The tub is being filled by the spout at a rate of 5 gallons/min, while at the same time water is being drained out of the tub through the hole at a rate of 10% of the bathtub water volume/min. How much water is in the bathtub?

Solution: This is a problem which can be solved easily if it is looked at as a system with a definite answer in steady state, where there is a negative feedback between the amount of water in the bathtub and the amount draining from the bathtub. If the bathtub has more than 50 gallons, the amount of water in the bathtub will be decreasing; if the bathtub has less than 50 gallons, the amount of water in the bathtub will be increasing. Therefore, the bathtub has exactly 50 gallons.

What has this to do with lasers, you ask? The rate of photon loss due to the cavity is constant (like the spout in the bathtub) and the rate of photon addition has to do with the gain that is dependent on the carrier density (like the leak). This is perhaps a loose analogy, but is a vivid image.

5.4 Absorption Loss

In reality, a few more parameters are necessary to make this model really useful. First, the cavity defined in Fig. 5.1 has a certain absorption loss associated with it. The light in the cavity experiences optical gain as it travels back and forth within the cavity, but it is also absorbed by mechanisms that do not depend on the carrier injection. Let us first include this absorption parameter as a phenomenological part of the cavity model, and then briefly discuss the mechanisms for absorption.

Including an absorption loss in the cavity leads to the following round trip expression for the gain,

$$\begin{aligned} 1 &= R_1 R_2 \exp(2gL) \exp(-2\alpha L) \\ g_{\text{cav}} &= \frac{1}{2L} \ln\left(\frac{1}{R_1 R_2}\right) + \alpha = A'(J_{\text{th}} - J_{\text{tr}}) \end{aligned} \tag{5.6}$$

which defines the lasing gain in terms of cavity parameters and absorption loss.

The first term, $\frac{1}{2L}\ln\left(\frac{1}{R_1R_2}\right)$ above, in Eq. 5.6 is called the distributed mirror loss. This term represents the photons 'lost' through the mirrors, as if that mirror loss is a lumped parameter over the entire laser length. The absorption loss, similarly, represents the optical loss due to absorption of photons through free carriers, scattering off the edges of ridges, or other means.

This absorption loss is *not* the optical absorption across the band gap—that absorption becomes gain as the material is pumped into population inversion. There are several mechanisms that are not carrier-density-dependent which induce optical absorption. Let us briefly discuss them.

5.4.1 Band-to-Band and Free Carrier Absorption

The most significant additional absorption factor in laser design is called 'free-carrier absorption.' This mechanism is illustrated below in Fig. 5.3 and is contrasted to band-to-band absorption.

Values of the band-to-band absorption coefficient are given by the expression in Eq. 4.37 and depend on the quasi-Fermi levels. (Negative gain, with quasi-Fermi level splits below the band gap, mean absorption rather than gain.) For lasers pumped into population inversion, there is band-to-band gain, not absorption; the gain term in Eq. 5.6 is due to band-to-band transitions.

A sub-class of band-to-band absorption is called *excitonic* absorption, often seen at very low temperatures or sometimes in very pure semiconductors and quantum wells at higher temperatures. An exciton is an electron-hole pair; at low temperatures, the electron and hole form a Coulombic attachment which lowers the energy of them both. This bound electron-hole pair is an exciton; when absorbed by a

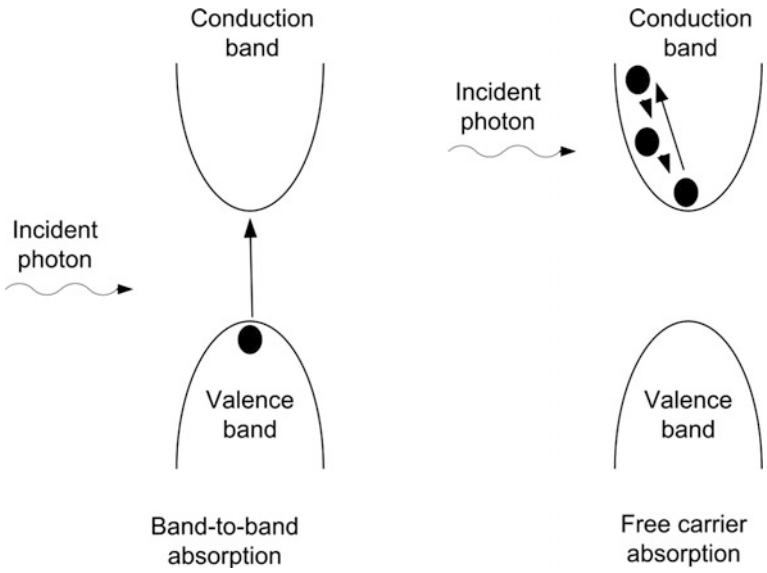


Fig. 5.3 **a** Band-to-band and **b** free carrier absorption. A photon is absorbed by a carrier (electron or hole) but instead of promoting an electron from the valence band to the conduction band (left), it promotes a carrier from the bottom of its band up to the top. The carrier (say an electron) then loses energy by interaction with other electrons and the lattice and relaxes back to the bottom of the band

photon, this exciton is removed. Extra absorption peaks seen at a semiconductor band edge are due to excitonic absorption.

Free-carrier absorption is a loss factor in lasers and part of the α term in Eq. 5.6. The mechanism for it is given as follows. A photon is incident on a semiconductor and excites a carrier (electron or hole). This electron or hole is promoted higher in its own band. After being excited, the carrier relaxes back down to its equilibrium position in the band through interaction with the lattice and with other carriers. This process is dependent on the doping density—the higher the doping density, the more likely this absorption process will take place. For this reason, the separate confining region around the quantum well is usually kept undoped. Quantitatively, the free carrier absorption is given as a function of doping density by the expression

$$\alpha_{\text{freecarrier}} = \frac{nq^2\lambda^2}{4\pi^2 m n_r c^3 \varepsilon_0}\frac{1}{\tau} \tag{5.7}$$

where n is the free carrier density (or doping density), λ is the wavelength, m the carrier mass, and τ is a 'scattering time' associated with the relaxation time of the carriers once they are excited. Because of the wavelength dependence, relatively low energy (longer wavelength) photons in more highly doped areas experience more free carrier absorption.

Devices designed for high power operation go through special efforts to keep this absorption value low—for example, pump lasers designed for several hundred mW typically have absorption losses in the range 2–5/cm. High speed modulated devices for telecommunications have numbers closers to 20/cm.

Because this process depends on the density of carriers in the region near the semiconductor, typically the separate confining heterostructure region is kept lightly doped to reduce absorption losses. However, like many things, this is a trade-off—some positive effects of increased doping are better conductivity, and hence, lower heat dissipation. In addition, increased p-doping in the active region can lead to better modulation performance.

5.4.2 Band-to-Impurity Absorption

As a matter of completeness, we observe that light can generally be absorbed wherever a carrier can be absorbed and induced to transition from one energy state to another. For example, impurities in a semiconductor, which trap carriers, can also serve as absorption sites, and there is often low energy absorption from impurities to conductor or valence bands (or sometimes, between bands, such as between the heavy and light hole bands). These mechanisms are not very important in lasers—in general, the absorption energy is much lower than the lasing energy (for standard telecommunication lasers), and there are few impurities in good lasing material. This mechanism is pictured in Fig. 5.4.

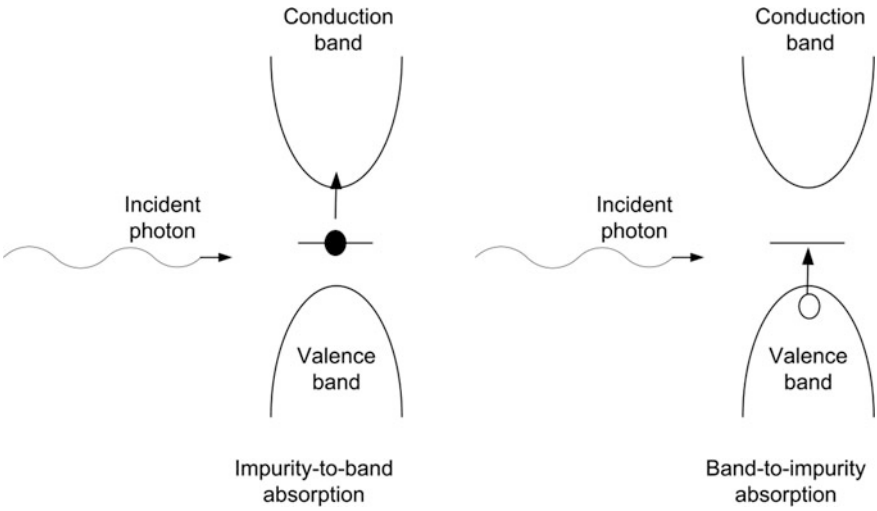


Fig. 5.4 Impurity to band and band to impurity absorption, illustrated. The horizontal line represents a defect state in the middle of the band gap. Typically lasers have few defects or impurities, and in addition, this mechanism is typically for much lower than band-gap energy photons

5.5 Rate Equation Models

One of the most useful and powerful tools to understanding laser operation is the rate equations. The idea is simple and best illustrated as we work through it. Figure 5.5 shows a schematic picture of a laser cavity, which contains a certain carrier density n and a photon number S. There are a number of things going on:

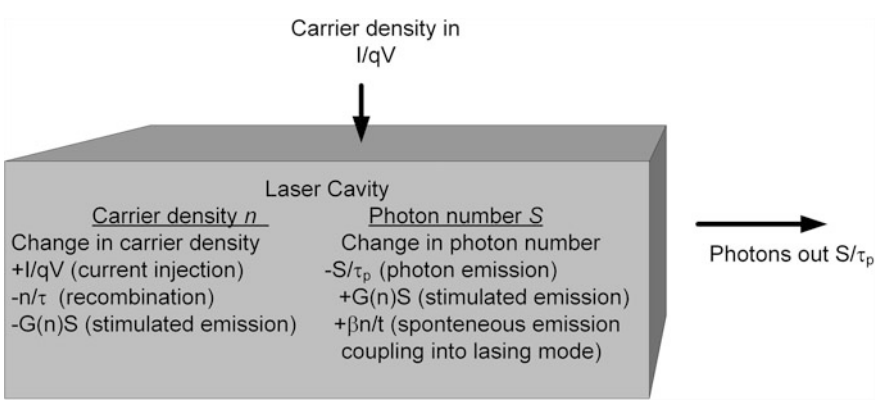


Fig. 5.5 Laser cavity, illustrating the processes which can change both photon number and carrier number

current is being injected, photons are coming out, and inside, carriers are being converted to photons through the mechanism of stimulated emission and spontaneous emission.

In the diagram, I is injected current, V is the carrier volume, q is electronic charge for each carrier, τ is carrier lifetime (which includes both radiative and non-radiative processes), and $G(n)$ is the gain as a function of carrier density (e.g., see Eq. 5.4).

All of these processes can change the carrier density and photon number in the cavity. We can write down a simple expression for all processes and set that quantity equal to the total rate-of-change in photon number or carrier density in the cavity. The expressions, and the mechanisms behind each term, are shown in Eqs. 5.8a and 5.8b.

$$\frac{dn}{dt} = \frac{I}{qV} - \frac{n}{\tau} - G(n)S \tag{5.8a}$$

Rate of change of carrier density

Injected current

Recombination (largely spontaneous emission)

Stimulated emission

$$\frac{dS}{dt} = S\left(G(n) - \frac{1}{\tau_p}\right) + \frac{\beta n}{\tau_r} \tag{5.8b}$$

Rate of change of photon density

Stimulated emission

Photon emission from cavity

Photon emission coupling into lasing mode

The first term on the right of Eq. 5.8a represents current injection. This current, in carriers/sec, is confined to some sort of volume V (the quantum well region) and exists for a carrier lifetime τ (and as well, being measured in Coulombs, means that it has a conversion factor from coulombs to carriers of q.) The second term represents the decay of carriers through natural recombination processes (including, but not limited to, radiative recombination). As each carrier exists for only τ seconds, the rate of density decline is n/τ.

The third term expresses the fact that for every photon generated through stimulated emission, carriers are lost. The expression $G(n)$ is a convenient expression which captures both the correct units and the dependence of gain on carrier density. Other forms, other than Eq. 5.6, are also used. The expression $G(n)$ here represents the modal gain (or the gain experienced by the optical mode)

rather than material gain (which would be the gain experienced by the optical mode if all the light were confined completely to the gain region). The left-hand side of Fig. 1.5 illustrates that the optical mode usually only fractionally overlaps the quantum well region; Chap. 7 will discuss this in more detail.

Equation 5.8b is a rate equation for the number of photons in the lasing mode (there are typically also many other additional photons at other wavelengths being created through spontaneous emission). They increase through stimulated emission ($G(n)S$) and are lost through the cavity facets and through absorption (S/τ_p). Both of these factors are proportional to the photon density S, and so S is factored in the parenthesized expression above.

A small fraction β of the photons created through spontaneous radiative recombination n/τ_r is at the correct wavelength, and in phase with, the lasing mode. These photons are said to 'couple' into the lasing mode. Typically this is not important except for mathematically kickstarting stimulated emission, which requires an initial, small, density of photons. The fraction of photons coupled into that mode, β, is of the order of 10^{-5}.

5.5.1 Carrier Lifetime

This is an appropriate place to talk for a moment about one of the time constants in the rate equations, the carrier lifetime τ. The spontaneous emission carrier lifetime is the typical amount of time that a carrier exists in the active region before it recombines and vanishes. The time constant is due to all mechanisms except for carrier depletion through stimulated emission.

There are actually several different ways a carrier can recombine, illustrated in Fig. 5.6. The most familiar is a direct bimolecular radiative recombination as shown in Fig. 5.6(left side). An electron recombines with a hole, and the energy taken up by an emitted photon. If there are defects in a material, the electron (or hole) can fall into the defect, where it is eventually eliminated when a carrier of the opposite species falls into the defect and renders it neutral again. In this case, the energy is taken up by phonons. This is called Shockley-Read-Hall recombination, or trap-based recombination, and is illustrated by Fig. 5.6(middle).

Finally, the mechanism of Auger recombination is illustrated in Fig. 5.6(right). In this mechanism, an electron and a hole recombine, but instead of emitting a photon, the energy is transferred to another carrier. That third carrier is kicked up higher in energy and serves to heat up the carrier distribution. Auger recombination as pictured here uses two electrons and one hole; however, it can take place with two holes and one electron and can involve transitions between bands (such as the heavy hole and light hole band). The essential feature is that it is a non-radiative method that requires three carriers and transfers the recombination energy to the third carrier instead of emitting a photon.

The relative importance of these three rates of recombination can be seen by writing the total spontaneous recombination rate R_{sp} (in s^{-1} cm^{-3}) as

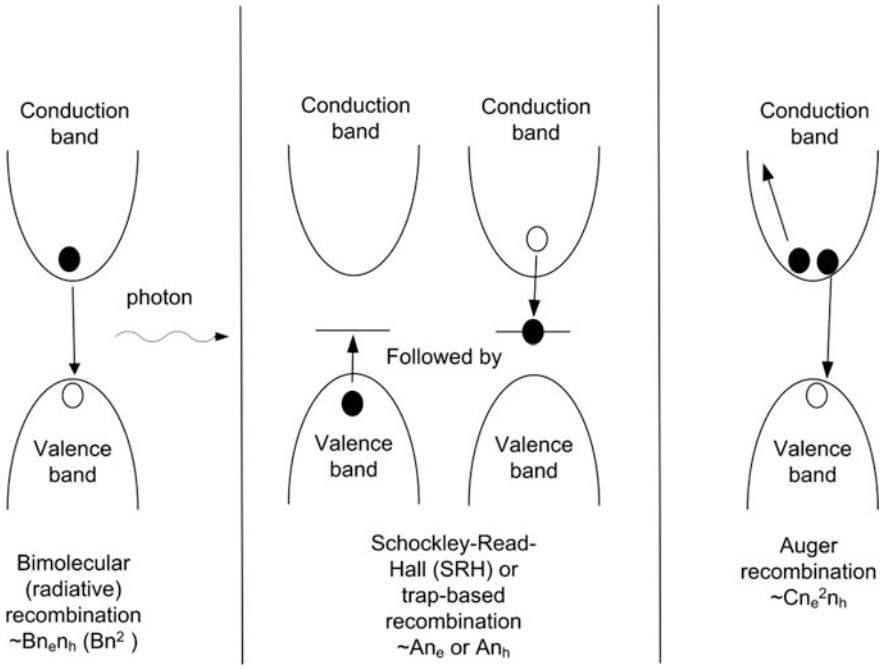


Fig. 5.6 Mechanisms of carrier recombination: bimolecular, trap-based, and Auger

$$R_{sp} = An + Bn^2 + Cn^3, \tag{5.9}$$

where An represents the rate of trap-related recombination, Bn^2 is the rate of bimolecular (radiative) recombination, and Cn^3 is the rate of Auger recombination. If the recombination rate is higher, the carrier lifetime is reduced. The impact of carrier lifetime on laser threshold current, for example, will be seen in Eq. 5.15, forthcoming. Here, we do not distinguish between electrons n_e or holes n_h; generally, (particularly in undoped laser active regions) they are both about the same and denoted by n.

Good lasers typically have very low defect densities, so the trap-based recombination term is often negligible. The dominant term for shorter wavelength devices (such as 980 nm) is bimolecular recombination. For longer wavelength (lower energy and band gap) devices, Auger recombination is more significant, and, as seen by Eq. 5.9, at higher carrier density, Auger is also more significant. In terms of recombination rate R_{sp}, recombination time τ can be written as

$$\tau = n/R_{sp} \tag{5.10}$$

In general, the carrier lifetime τ in laser rate equations is about 1 ns.

Having defined and discussed τ, let us look further into the rate equation model.

5.5.2 Consequences in Steady State

For in a laser in steady state, all of these observable quantities—*n, S*, and *I*—are not changing with time. It does not matter if we look at the laser now or twenty minutes from now; it will look the same. Let us look at what these rate equations tell us when the rates of change, d*n*/d*t* and d*s*/d*t*, are zero.

Let us look at the second expression first, in steady state.

$$0 = S\left(G(n) - \frac{1}{\tau_p}\right) + \frac{\beta n}{\tau_r} \approx S\left(G(n) - \frac{1}{\tau_p}\right) \tag{5.11}$$

We will neglect the $\beta n/\tau_r$ term—it is relatively small compared to the density of photons created due to stimulated emission. The equation then says that either $S = 0$ (low photon density), or the gain $G(n) = 1/\tau_p$. (we will discuss the question of the units of gain in a moment—here, they are clearly in units of s^{-1}.)

The gain $G(n)$ obviously depends on n, while the photon lifetime in the cavity depends only on things like the facet coating and optical absorption and not on n. Therefore, the first, very important observation is that the gain $G(n)$ is clamped at the threshold carrier density n_{th} to a value $G(n_{th})$ set by the laser cavity and does not increase further with increased carrier injection. This is the same conclusion, restated, that was obtained in Sect. 5.3.

Hence, the actual value of the lasing gain is set fundamentally by the cavity, not by the mechanics of the gain region. By far the most effective way to alter the lasing gain, and consequently, parameters like threshold current, is to change cavity characteristics including the length and threshold coating. The properties of the active region substantially set the threshold current density n_{th}.

Below this 'threshold' carrier density, the photon density is approximately zero. At n_{th}, the gain is clamped by the cavity properties.

Let us take a look at Eq. 5.8a in light of this observation.

$$\begin{aligned} 0 &= \frac{I}{qV} - \frac{n}{\tau} - G(n)S = \frac{I}{qV} - \frac{n}{\tau} \quad \text{for } n < n_{th}\ (S = 0) \\ 0 &= \frac{I}{qV} - \frac{n}{\tau} - G(n)S = \frac{I}{qV} - \frac{n_{th}}{\tau} - G(n_{th})S \quad \text{for } n = n_{th}\ (S > 0) \end{aligned} \tag{5.12}$$

Equation 5.12 above, for n below and up to threshold carrier density (when the photon density is 0) simply says that injected current linearly increases the carrier density as

$$n = \frac{I\tau}{qV} \tag{5.13}$$

Every injected carrier exists for a characteristic time τ, occupies a volume V, and has charge q converting current to carriers. Equation 5.13 can almost be written

down directly from a common-sense perspective. Typically, the lifetime τ (including recombination processes except stimulated emission) is about 1 ns.

If $n = n_{th}$ (remember, we have concluded that n cannot be greater than n_{th}) we can write Eq. 5.12 as

$$S = \frac{1}{G(n_{th})}(I - I_{th}), \tag{5.14}$$

where I_{th} is the threshold current is defined from Eq. 5.13, where $n = n_{th}$ as

$$I_{th} = \frac{qVn_{th}}{\tau} \tag{5.15}$$

Equations 5.12 and 5.14 predict the easily-observed laser properties in Fig. 5.7. Below a certain threshold current I_{th}, there is very little light out. The current injected serves to increase the carrier density. Above the threshold current density, the carrier density is clamped and further increases in current increase the photon density.

Just as the photon density (and the light out of the cavity) changes qualitatively at the threshold current, the electrical properties also change qualitatively (but subtly) at threshold. This will be discussed in Chap. 6.

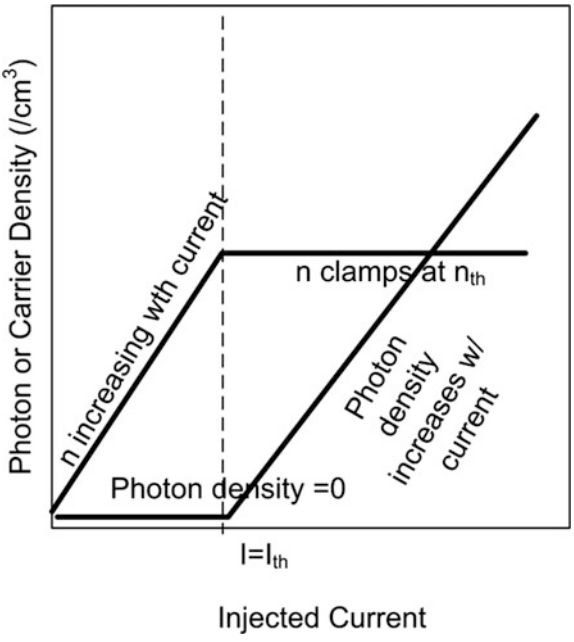


Fig. 5.7 Predictions of the rate equations with respect to carrier density n and photon density S. Below threshold, the current density is clamped with a nominal photon density due only to spontaneous emission; above threshold, the carrier density is clamped, and the photon density increases linearly with injected current

5.5.3 Units of Gain and Photon Lifetime

In Chap. 4, and photon lifetime at the beginning of this chapter, we wrote down an expression for gain in terms of cm^{-1} as defined by its exponential dependence on length, $P = P_0 \exp(gx)$. In the rate equation model, it is clear that $G(n)S$ has to have units of s^{-1}. Which is correct?

The answer is both. Gain in cm^{-1} can be converted to gain in s^{-1} by using as conversion factor the velocity of light, as shown below.

$$g[\mathrm{cm}^{-1}] = g[\mathrm{s}^{-1}]\frac{c}{n} \tag{5.16}$$

where c/n is the group velocity, and v_g is the velocity of light in the medium.

We also note that we have very casually written gain as proportional to current, current density, carrier density, and carrier number, and with units of either cm^{-1} or s^{-1}. In the context, in which we use these simple gain models, these are all basically correct. The prefactor A is picked to give the correct units for whatever proportionality we find currently convenient.

Example: Estimate the photon lifetime in a 300 μm-long laser device with uncoated facets and an index of 3.5.

Solution: The calculated gain point is given by Eq. 5.6, and is 40/cm.

Dividing by c/n gives a value of $1/\tau_p$ of 3.3×10^{11}/s, or a time constant τ_p = 3 ps.

This small ps photon lifetime is fundamentally the reason that semiconductor lasers can be rapidly modulated. When we rapidly change the current going into the device, the photon density can also rapidly change.

In contrast, modulated light-emitting diodes are driven by spontaneous emission, and the light from those devices is proportional to n/τ, where τ is the carrier lifetime (typically in ns). Because laser light is limited largely by photon lifetime of ps, while light from a light-emitting diode is limited by carrier lifetime of ns, lasers can be modulated at Gb/s speeds which are much faster than diode speeds. This is fundamentally why optical communication requires lasers.

5.5.4 Slope Efficiency

Figure 5.8a shows the most basic of all laser measurements—a light-current, or L-I, curve. A current source injects a precise amount of current into the laser bar, and an optical detector in from the bar measures the amount of light L (in Watts, W) out of

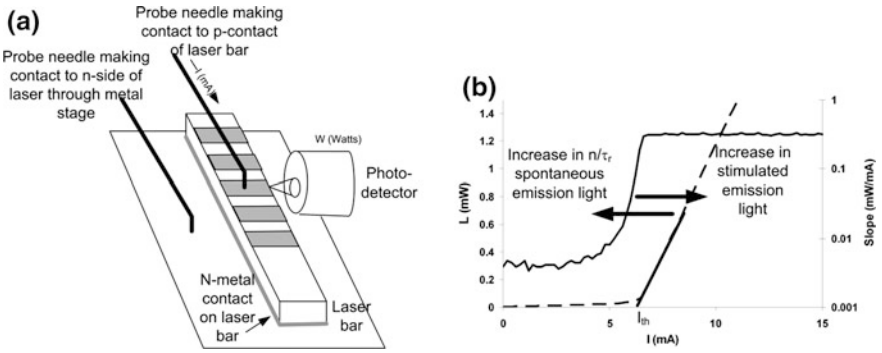


Fig. 5.8 **a** Measurement setup for a laser bar and **b** the L-I measurement of the device

the device. Figure 5.8b shows two items of data derived from the measurement—first, the light out as a function of the current in, and second, the derivative (dL/dI) or slope, in W/A, versus the current in.

Notice how exactly this behavior matches the predictions of the rate equations. There is an abrupt increase in the amount of light out, at a particular threshold current I_{th}, proportional to the current. The slope of that proportionality (in Watts out/Amps in) is usually called the slope efficiency (abbreviated as SE) and is something that has a minimum specification in a commercial device. Generally, the higher the slope efficiency, the better we want to extract as much light per given injected current as possible.

There are several definitions of threshold current from a measured L-I curve. The most common is the current extrapolated back to the point where the light is zero, or about 6 mA in Fig. 5.8b. Other definitions are the point of maximum slope or the point where the slope changes.

Let us quantify the slope efficiency in terms of the cavity parameters R_1, R_2, and α. Suppose an amount of current I is injected into the device, and of that current, a fraction η_i (the internal quantum efficiency) is converted into photons. Those photons in the laser cavity then are either re-absorbed (represented by the loss α) or emitted out of one of the facets (represented by the distributed optical loss, $1/2L \ln(1/R_1R_2)$ (in this expression, L is cavity length). The latter term, while it represents 'loss' in terms of the gain needed, actually represents photons exciting the cavity and is desirable.

The ratio of external quantum efficiency (η_e) in photons out/carriers in to internal quantum efficiency, in terms of the photons exciting the cavity and the photons absorbed within the cavity, is given by the expression

$$\eta_e = \frac{\eta_i \frac{1}{2L} \ln\left(\frac{1}{R_1 R_2}\right)}{\frac{1}{2L} \ln\left(\frac{1}{R_1 R_2}\right) + \alpha} \tag{5.17}$$

The ratio of external conversion efficiency to internal conversion efficiency is equal to the ratio of distributed optical loss to total loss.

Both η_i and η_e are in terms of photons/carrier, while the quantity that is measured (in the measurement pictured in Fig. 5.8a) is the slope efficiency in W/A. Each photon of wavelength λ carries an energy of 1.24 eV μm/λ, and the conversion between eV and V is the electron charge q. The relationship between slope efficiency SE in W/A and η_e is then

$$\mathrm{SE(W/A)} = \frac{1.24}{\lambda(\mu\mathrm{m})}\eta_e(\mathrm{photons/carrier}) \tag{5.18}$$

Usually, slope efficiency is typically measured out of only one facet. If the facet reflectivity is the same, then that number can be doubled to determine the total Watts/A emitted from the device. When the facet reflectivity is different, as is usually the case, additional analysis is needed.

Equation 5.17 is an expression that can be used to determine both the internal loss α and the internal quantum efficiency of a laser material, based on a set of measurements of devices that are of varying length but are otherwise identical. If the equation is re-written as

$$\frac{1}{\eta_e} = \frac{1}{\eta_i}\left(1 + \frac{2L\alpha}{\ln\left(\frac{1}{R_1R_2}\right)}\right) \tag{5.19}$$

it is clear that the slope increases as the device gets shorter and that the extrapolated value (where the cavity length $L = 0$) will give the internal quantum efficiency η_i. This fraction of injected carriers that are converted to photons is an important figure of merit for the material and is typically of the order 80–100%. This process also illustrates the methodology behind much of laser analysis—through fairly simple models, material constants are related to measurements.

5.6 Facet-Coated Devices

In most applications of semiconductor edge-emitting lasers, the facet reflectivities of the two facets are not equal. In edge-emitting Fabry-Perot lasers, the mirrors are first formed by physical cleave of the wafer (Fig. 5.9). The wafers are scribed (scratched) on an edge with a diamond-tipped tool, and then broken; the break propagates along the crystal planes forming a perfect dielectric mirror between the semiconductor and air. As formed, these mirrors are symmetric, and so half of the light would exit one side of the cavity and half the other. It is important when doing this to align the scribe and cleave marks with the plane of the wafer which is being cleaved.

While perfectly acceptable as a textbook example, for commercial purposes, it is desirable that most of the light exit one facet to be coupled into an optical fiber. Hence, the facets are usually coated with dielectric coatings in order to modify the

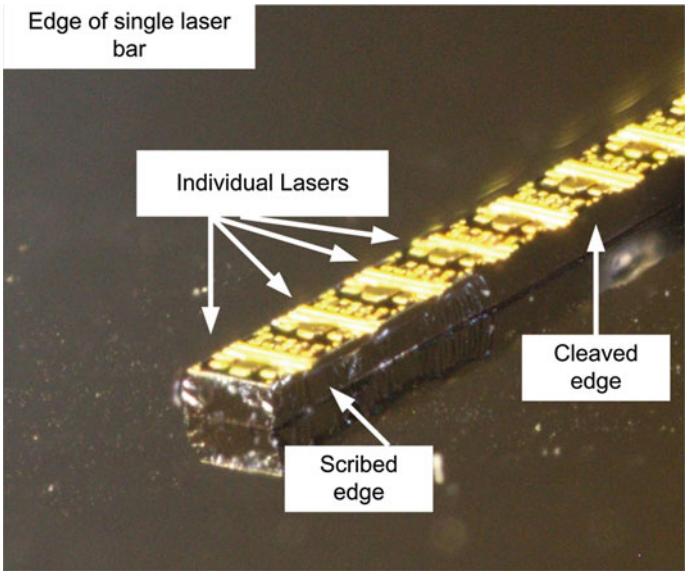


Fig. 5.9 Laser bar, showing (left) a scribed edge, where the break was started, and mirror-flat cleaved edge, which creates the mirror for the laser cavity. Where it was scribed, the devices do not lase and are discarded. Photo credit J. Pitarresi

reflectivity. A typical design for a Fabry-Perot laser has a rear facet reflectivity of about ~70%, and a front facet reflectivity of ~10%. Most of the light exits the laser from the front facet with a small amount exiting the rear facet. The rear facet light is often coupled to a monitor photodiode in the package, to enable active control of the output laser power. Typical Fabry-Perot laser coatings are shown in Fig. 5.10.

These coated facets are an excellent way to control the laser properties. From Eq. 5.6, it is clear that required cavity gain decreases as the facet reflectivity increases. Hence, the threshold current required can be reduced by increasing the facet coating reflectivity.

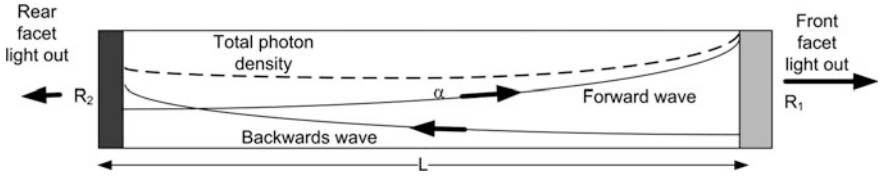


Fig. 5.10 Typical telecommunications Fabry-Perot laser, with one side HR coated to 70% reflectivity, and the other side LR coated to 10% reflectivity. Notice the asymmetry, with most of the light near the front facet

Example: Calculate the value of the lasing gain point of the cavity pictured in Fig. 5.5, where $R_1 = 0.1$ and $R_2 = 0.7$. Compare it the value of the lasing gain point of the cavity if the facets were uncoated, with $R_1 = R_2 = 0.35$. Neglect absorption loss.

Solution: From Eq. 5.6, with $L = 500$ μm, the gain point is

$$53 = \frac{1}{2(0.05)} \ln\left(\frac{1}{(0.7)(0.1)}\right)$$

If the facets were both uncoated, with reflectivity of 0.3, the gain point would be 72/cm.

If the reflectivity of the two facets are not equal (and they usually are not), then the slope efficiency out of the two facets is also different. The term asymmetry means the ratio of the slope efficiency out of one facet SE_1 over the slope efficiency out of the other facet SE_2, and for Fabry Perot lasers is given

$$\frac{SE_1}{SE_2} = \frac{R_1^{-1/2} - R_1^{1/2}}{R_2^{-1/2} - R_2^{1/2}} \tag{5.20}$$

Tailoring the slope efficiency is a useful and powerful way to affect the performance of the laser.

Example: A Fabry-Perot 1.48 μm laser has a low reflectivity (LR)/high reflectivity (HR) pair of facet coatings with reflectivity $R_1 = 0.1$ and $R_2 = 0.7$, respectively, and is intended to have a fiber coupled to the LR side. The internal quantum efficiency is 0.8, and the absorption loss is 15/cm. For a cavity length of 400 μm, calculate the slope efficiency in W/A out of the front facet.

Solution: The total slope efficiency in photons/carrier is calculated using Eq. 5.19 to be 0.55.

$$0.55 = \frac{0.8 * \left(\frac{1}{2(0.04)} \ln\left(\frac{1}{(0.7)(0.1)}\right)\right)}{\frac{1}{2(0.04)} \ln\left(\frac{1}{(0.7)(0.1)}\right) + 15}$$

According to Eq. 5.20, the ratio of the slope out the front to slope out the back facet is

$$7.9 = \frac{0.1^{-0.5} - 0.1^{0.5}}{0.7^{-0.5} - 0.7^{0.5}}$$

Hence, the slope efficiency in photons/carrier out the front facet is

$$0.49 = \frac{7.9}{8.9} 0.55$$

And in W/A,

$$0.41 = 0.49 \frac{1.24}{1.48}$$

Later in Chap. 8, we will extensively discuss another type of device called a distributed feedback (DFB) laser. Those lasers are also coated, but in those devices the equations for relative power given in this chapter do not apply.

5.7 A Complete DC Analysis

Fundamentally, laser characteristics are limited first by the material and then affected by the structure. The kinds of samples used for material analysis are almost always 'broad-area' samples, tested with pulsed current sources. These types of samples and testing methods are used to avoid non-idealities associated with the waveguide that we are trying to measure material properties and with heating effects. (Laser devices exhibit significant heating effects at higher current.)

Figure 5.11 illustrates the difference between broad-area and single-mode (ridge waveguide) devices.

Several different devices are measured at each length because there is significant variation from device-to-device.

The two key equations in this sort of analysis are Eqs. 5.6 and 5.19. Shown below is an example of the complete set of data acquired from devices of various lengths and the analysis of material and device properties.

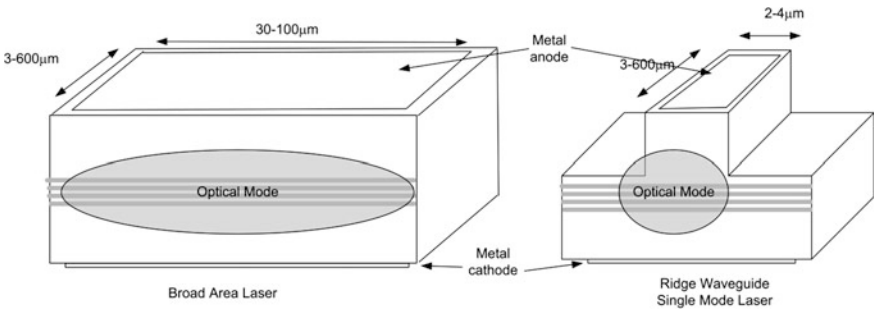


Fig. 5.11 Left, broad area, and right, ridge waveguide devices. Ridge waveguide support single transverse mode operation and are used for communication, while broad-area devices are used for material characterization as details of the ridge, and resistance, matter much less

Example: The set of data in Table 5.1 is obtained on broad-area laser devices which have a lasing wavelength of 1.31 μm. Find the transparency current of this material, the. absorption loss, and the internal quantum efficiency.

Table 5.1 A set of data obtained from a few different laser samples each with a 30 μm stripe width and uncoated facets

Sample #	Sample length (μm)	I_{th} (mA)	SE (measured from one facet) (W/A)	J_{th} (I_{th}/ length * 30 μm)	SE (two facets, in photons/carrier)
		Measured quantities		Calculated quantities	
1	500	217	0.14	*1447*	*0.30*
2	500	217	0.13	*1447*	*0.27*
3	500	217	0.18	*1447*	*0.34*
4	750	259	0.09	*1151*	*0.19*
5	750	269	0.09	*1187*	*0.23*
6	750	258	0.11	*1147*	*0.21*
7	1000	286	9.1×10^{-2}	*953*	*0.19*
8	1000	294	9.2×10^{-2}	*980*	*0.19*
9	1000	297	8.0×10^{-2}	*990*	*0.17*

The columns at left, I_{th} (mA) and SE (W/A) are directly measured quantities; the columns at right, J_{th} and SE (photons/carrier)

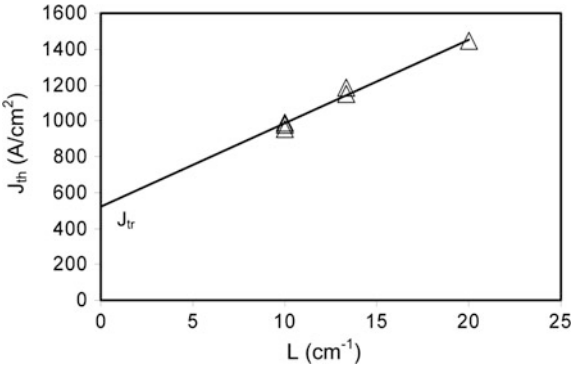


Fig. P5.12 Threshold current density versus 1/*L* for a set of lasers, showing J_{th} about 500 A/cm^2

Solution: The straightforward process is illustrated by example below. The theoretical model is provided by Eqs. 5.6 and 5.19. First, the current density is calculated by simply dividing by the area. The measured output efficiency is evaluated by multiplying by two (in this case, where the facets are identically uncoated) and by $\lambda/1.24$ eV μm. These values are plotted in the last two columns of Table 5.1.

To determine transparency current, the threshold current density is plotted versus $1/L$ according to Eq. 5.6. The result is shown in Fig. P5.12. The value extrapolated as L tends to infinity is the transparency current density which is the minimum current density

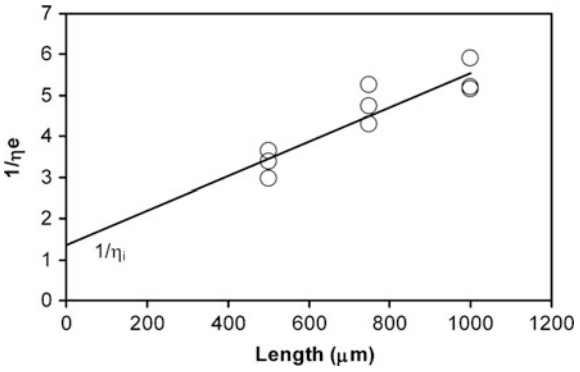


Fig. P5.13 External quantum efficiency versus device length *L*. The intercept gives the internal quantum efficiency, while the absorption loss can be obtained from the slope

required to lase in this material. This number is often used as a figure of merit for the material.

The efficiency versus length can be plotted according to Eq. 5.19. This equation shows the relative effect of mirror loss versus absorption loss. As the cavity length goes to zero, the only effective loss is the mirror loss, and the ratio of carriers into photons out gives the internal quantum efficiency (typically >0.60). Below, $1/\eta_e$ (external quantum efficiency) is plotted as a function of L to show extracted internal quantum efficiency of about 0.74.

The slope plotted in Fig. P5.13 gives the absorption loss α. (If this value is measured in a broad-area device, it can be different than that seen in a ridge waveguide, due to the scattering from the ridge.)

The best-fit equation for $1/\eta_e$ versus L in Fig. P5.13 is

$$\frac{1}{\eta} = 0.0042L + 1.36$$

Comparing with Eq. 5.19, $0.0042 = 2\alpha/\eta_i * 1/\ln(R_1R_2)$, and with known facet reflectivities $R_1 = R_2 = 0.3$, and extracted value of η_i of 0.74, gives a value for α of 3.74×10^{-3} μm^{-1}, or 37 cm^{-1}.

5.8 Summary and Learning Points

In this chapter, we related the fundamental internal properties of semiconductor quantum wells to the input and output parameters of a device.

A. The reflectivity of as-cleaved semiconductor facets is given by the index of the material and air and is typically about 0.30.
B. Lasers operate in a steady-state condition of unity round trip gain where for a constant current input (or any input excitation level) the photon density in the cavity and exiting the cavity is stable.
C. A simple but useful model of the gain represents it as proportional to the carrier density minus a transparency carrier density. The transparency current density

is a structure and material constant that sets the minimum carrier density at which the material can lase.

D. In addition to the gain and loss associated with the active region, there is absorption loss associated with absorption of the optical mode in the doped cladding layers. There is also optical scattering from the waveguide. These additional loss terms affect the efficiency and threshold current of the device.
E. The gain point of a Fabry-Perot optical cavity is set by the absorption losses and the facet reflectivity.
F. Threshold current and slope efficiency of a given device are affected by facet reflectivity. Commercial devices typically have their facets coated to cause more light to exit the primary end.
G. By evaluation of threshold current density as a function of length, a material/structure parameter called transparency current density can be measured. This sets the minimum threshold current density obtainable for a very long device and is used as a figure of merit for laser structures.
H. Rate equation models are used to relate injection current, carrier density and photon density and predict the DC characteristics of threshold and linear L-I slope that are observed.
I. Gain can be expressed in cm^{-1} (as appropriated for the optical loss equation) or in s^{-1} (as in the rate equation) and are appropriately related by the speed of light in the medium.
J. The short photon lifetime in a semiconductor laser cavity is fundamentally the reason that they can be modulated very rapidly.
K. The total slope efficiency is given by the ratio of optical loss to total loss.
L. By analysis of DC characteristics of threshold current density and slope efficiency versus length, cavity and material/structure parameters such as internal quantum efficiency, absorption loss, and transparency can be extracted. These numbers are often used as figures of merit for a structure or material.

5.9 Questions

Q5.1. True or False. The amplitude and power reflectivity at the interface of a semiconductor facet and air increases as the dielectric constant of the semiconductor increases.
Q5.2. Would the power coming out of an uncoated semiconductor laser increase if it were tested in water instead of air?
Q5.3. True or False. Every photon that is created by recombination involves the removal of an electron and a hole.
Q5.4. What physical properties of a cavity determine the steady-state DC lasing gain?
Q5.5. What happens to the cavity gain g and threshold current I_{th} when the reflectivity of the facets R_1 and R_2 is increased?

Q5.6. What happens to the cavity gain g as the cavity length increases? What happens to the threshold current I_{th}?
Q5.7. What phenomena determine absorption loss? Is absorption loss minimized or maximized in manufacturing real semiconductor lasers?
Q5.8. What is the rate equation model for lasing (see Eq. 5.12 and describe the physical mechanism behind each term).
Q5.9. What is transparency current and how is it determined?
Q5.10. What is an L-I curve?
Q5.11. Define external and internal quantum efficiency. How are these properties measured?
Q5.12 Why are measurements for fundamental properties such as transparency current usually done with broad-area lasers and pulsed current?
Q5.13. What is slope efficiency?
Q5.14. What are typical values of the reflectivities of both facets of a Fabry-Perot semiconductor laser in order to allow most of the light to couple to an optical fiber attached to one facet?

5.10 Problems

P5.1 A semiconductor laser has a threshold current I_{th} of 20 mA with a carrier lifetime of 1 ns (due to Auger and bimolecular recombination) and an impurity density of $<10^{13}/cm^3$. Figure P5.14 gives the dependence of carrier lifetime on impurity density in this particular material.

(a) By what mechanism does increasing impurity density reduce the lifetime?
(b) If the laser had an impurity density of $10^{18}/cm^3$, what would its threshold current be?

P5.2 A laser designed to laser at 980 nm has an internal efficiency of 0.9, power reflectivity of 0.4 from both facets, a length of 300 μm, and internal absorption loss of $20/cm^{-1}$.

(a) What is the photon lifetime τ_p?
(b) What is the slope efficiency, measured out of one facet, measured in W/A?

P5.3. A laser active region has the following material properties:

η_i (internal quantum efficiency)	0.8
J_{tr} (transparency current density)	2000 A/cm^2

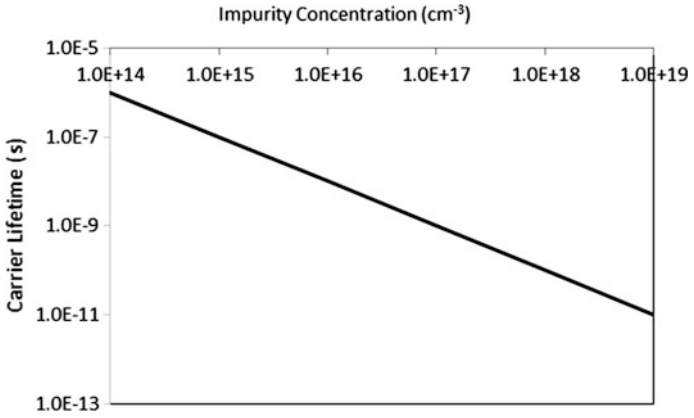


Fig. P5.14 Recombination lifetime versus impurity density for some semiconductor

A (differential gain)	0.02 (/cm * cm^2/A)
E_g (band gap)	0.946 eV.

In addition, the waveguide structure used is 1 μm wide and has an additional loss

α (absorption loss) 20/cm.

Design a laser with the following properties:

Front facet slope greater than 0.4 W/A
Rear facet slope at least 0.05 W/A
Length between 150 and 450 μm
Threshold current below 20 mA.
The actual design can be done with a spreadsheet, but for what you submit, please calculate explicitly the threshold current, and slope efficiency out of each facet as a function of your chosen parameters.

P5.4. Vertical cavity lasers use dielectric Bragg stacks as mirrors and can be made with extremely high reflectivity. The mirrors are typically circular, and the active area, instead of being set by the length times the ridge width, is set by area πr^2. The table below summarizes the length, ridge width, calculated active area, and reflectivity of a typical edge-emitting laser, as wells as the length, radius, and calculated active area for a VCSEL (Fig. P5.15).

(a) Calculate the mirror loss for the edge-emitting laser.

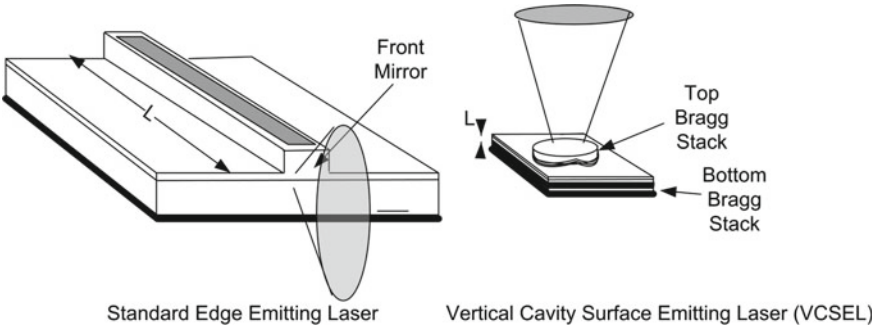
(b) Calculate the reflectivity for the VCSEL which will give it the same mirror loss as the edge-emitting laser.
(c) Assuming these cavities are crafted from the same gain region, and neglecting absorption, estimate the threshold current for the VCSEL.

P5.5. The rate equation model, above, predicts a threshold current where $n = n_{th}$ above which the light out is linearly proportional to current density n. This can be easily derived if we assume that $\beta n/\tau_r$ is negligible. However, spontaneous emission is observed below threshold, and light-emitting diodes operate completely through the means of spontaneous emission.

Derive the sub-threshold slope ratio of *S/J* in terms of other quantities in the rate equation for $n < n_{th}$.

P5.6. An uncoated laser has a facet active area of A, a modal index of n (which determines both reflectivity and mode speed), and a facet reflectivity of R. Assuming a uniform photon density in the optical cavity, determine an expression for photon density in the cavity in terms of power measured P (in W) out of the cavity facet.

P5.7. The 1 mm long device in the example of Sect. 5.7 has a threshold current of about 290 mA with uncoated facets. If the device was coated with facets >99% reflectivity (to reduce the facet reflectivity to negligible levels), what would its threshold current be?



Edge emitting laser properties	Surface emitting laser properties
L=300μm	L=1μm
R_1=R_2=0.3	R_1=R_2=?
Ridge width=1.5μm	Diameter=2μm
I_{th}=10mA	I_{th}=?
Active area=4.5x10^{-6}cm^2	Active area=3X10^{-8}cm^2

Fig. P5.15 Picture and properties of edge-emitting and surface-emitting laser devices. The shaded area represents the emitted optical mode

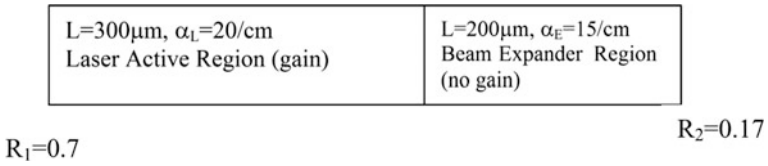


Fig. P5.16 Laser with a partial active cavity

P5.8. Fig. P5.16 shows a laser with a partial active cavity. In this structure, the part on the left is the active region with the quantum wells and gain; the part on the right is a 'beam expander,' which has no gain but is engineered to change the pattern of light out of the device to something that will better couple into optical fiber (glance ahead at Fig. 7.11!). As seen in Fig. 5.10, the general power distribution in a laser cavity is non-uniform. This problem involves modeling the cavity above to calculate the power distribution in this unusual cavity.

(a) Find the gain point g in the active region at which this structure will lase.
(b) Plot the forward-going, backward-going, and total power distribution in this cavity.
(c) Find the slope efficiency out the front facet in terms of photons out/total photons created.

Electrical Characteristics of Semiconductor Lasers

6

Some say the world will end in fire
Some say in ice…
—Robert Frost

Abstract

In this chapter, the electrical characteristics of semiconductor lasers are discussed. The basic operation of *p–n* junction diodes is reviewed, and the ways in which semiconductor lasers are and are not diodes will be enumerated.

6.1 Introduction

In the first several chapters of the book, we have talked about the general properties of lasers and then the specifics of semiconductor lasers. More or less, our analysis has started at the active region—the 'fire'—and the way that the electrons and holes create lasing photons. However, there is another important part of it, which is how the electrons and holes make their way to the active region in the first place. This part—the 'ice', if the reader will allow the poetic analogy to be strained more than GaAs grown on as Si substrate—is not unique to semiconductor lasers, but is nonetheless crucially important to them.

In this chapter, we will review semiconductor *p–n* and *p–i–n* junctions, and then we discuss ways in which lasers diverge from ideal *p–i–n* junctions. We will also discuss metal contacts to semiconductor lasers. We do expect the reader to have encountered *p–n* junctions before, and so our treatment is terse. More details can be found in many other textbooks on semiconductors.[1]

[1]For example, Streetman and Banerjee, *Solid State Electronic Devices*, Prentice Hall.

© Springer Nature Switzerland AG 2020

D. J. Klotzkin, *Introduction to Semiconductor Lasers for Optical Communications*,
https://doi.org/10.1007/978-3-030-24501-6_6

6.2 Basics of *p–n* Junctions

Semiconductor laser diodes consist of a *p*-doped region on one side, a generally undoped region of quantum wells and barriers in the center which is the 'active region' of the laser diode and an *n*-doped region on the other side. Electrons are injected from one side, and holes are injected from the other side. Both electrons and holes accumulate in the active region.

The objective is to derive the *p–n* junction diode equation. Because there is a lot of math to follow, and as a navigational aide, we illustrate the logical flow in Table 6.1. Then we will see how the derived expression applies to lasers.

The result of all this is to derive a general expression for the *I–V* curve across a *p–n* junction. The salient features are an exponential dependence of current on voltage and a reverse saturation current that depends on the features of the active region (doping, mobility, and lifetime).

6.2.1 Carrier Density as a Function of Fermi Level Position

The very first thing to introduce, or more appropriately, remind the reader of, is that the Fermi level, E_f, is fundamentally a measure of carrier density. The number of holes or electrons is given by the relatively complicated expression in Table 4.1, which includes the Fermi distribution function and the density of states function.

Table 6.1 Steps in deriving the diode current equation

Step	Sections
1. The use of Fermi levels to describe the population of a single *p*- or *n*-doped semiconductor is demonstrated	6.2.1
2. The band structure of an abrupt *p–n* junction in equilibrium is drawn	6.2.2
3. From the band structure, the space charge region and built-in voltage is derived	6.2.3, 6.2.4
4. From the relationship between space charge, and voltage, the width of the space charge region is derived	6.2.5
5. The same abrupt junction has a bias applied to it, splitting the Fermi level into twoFermi levelsquasi-Fermi levels (one for electrons and one for holes)	6.3
6. From the band structure picture, a rough picture of the charge density is sketched, assuming (as usual) an abrupt transition between the depletion region (with only space charge and no mobile charge) and the quasi-neutral region (with no net charge)	6.3.1
7. Assuming the excess charge is given by the minority carrier expression, an expression for excess minority carrier charge is derived, and from that, minority carrier diffusion currents	6.3.2
8. Finally, because current is continuous, the total current across the junction (neglecting recombination current in the depletion region) is equal to the sum of minority carrierdiffusion currents on each side of the junction	6.3.3

Table 6.2 Band gap, intrinsic carrier concentration, effective density of states, and relative refractive index of some common materials

Material	Band gap (eV)	n_i (cm^3)	N_C (cm^3)	N_v (cm^3)	ε_r ($\varepsilon_0 = 8.85 \times 10^{-12}$ F/m)
Si	1.12	1.45×10^{10}	2.8×10^{19}	1.0×10^{19}	11.7
GaAs	1.42	9×10^{6}	4.7×10^{17}	7×10^{18}	13.1
AlAs	2.16	10	1.5×10^{17}	1.9×10^{17}	10.1
InP	1.34	1.3×10^{7}	5.7×10^{17}	1.1×10^{19}	12.5

However, for bulk semiconductors in which the Fermi level is not too close to the conduction or valence band, there are two convenient simplifications.

First, the number of electrons and holes, n_0 and p_0, in equilibrium, can be written as

$$\begin{aligned} n_0 &= N_c \exp(-(E_c - E_{\text{Fermi}})/kT) \\ p_0 &= N_v \exp(-(E_{\text{Fermi}} - E_v)/kT) \end{aligned} \tag{6.1}$$

where E_c and E_v are the energy levels of the valence and conduction bands, respectively. The terms N_c and N_v are what are called the effective density of states of the conduction band and valence band, respectively. This simplification lumps all the states in the bands into one number, located exactly at the conduction band edge, and so, rather than the expression in Table 4.1, only a multiplication is needed. This number is about 10^{20}/cm^3 in Si and 10^{17}/cm^3 in GaAs. Particular values for different materials are in Table 6.2.

The product $n_0 p_0$ has the property,

$$\begin{aligned} n_0 p_0 &= N_c N_v \exp(-(E_c - E_{\text{Fermi}})/kT) \exp(-(E_{\text{Fermi}} - E_v)/kT) \\ &= N_c N_v \exp(-(E_c - E_v)/kT) \\ &= N_c N_v \exp\left(-E_g/kT\right) = n_i^2 \end{aligned} \tag{6.2}$$

and is a constant in equilibrium, independent of the Fermi level. The number n_i is called the intrinsic number of carriers and is a material property. In an undoped semiconductor, this represents the density of bonds which will be broken thermally and create holes and electrons.

In most semiconductors, the carriers are created by doping, and typically n_0 or p_0 is set by the density of donor atoms, N_D, or acceptor atoms, N_A. The dopant atoms are things which fit into the lattice but are either deficient in electrons (Group III dopants, like B or C) or have an extra electron (Group V dopants, like As).

The effect is to set the Fermi level not at the intrinsic Fermi level (E_i, in the middle of the band gap) but either near the conduction band, for *n*-doped semiconductors, or near the valence band, for *p*-doped semiconductors.

For the moment, let us look at a Si lattice. Equation 6.2 says that if n_0 is increased (say, to 10^{17}/cm^3, by doping Si to a 10^{17}/cm^3 level), then the equilibrium

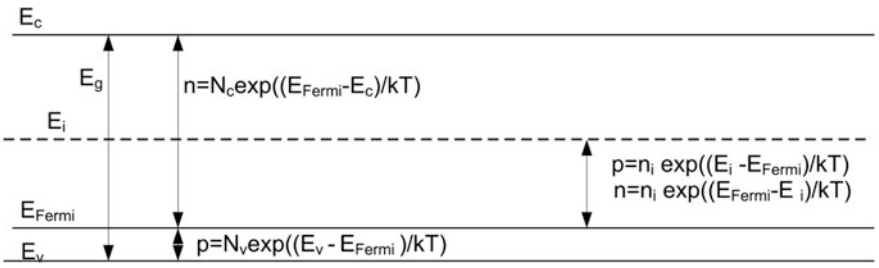


Fig. 6.1 Band structure of a *p*-doped semiconductor illustrating how the carrier concentrations can be referenced to the conduction band or to the intrinsic Fermi level

density of holes falls to $10^3/\text{cm}^3$. In an undoped semiconductor, mobile holes are created along with the mobile electrons, and so $n_0 = p_0$.

Equation 6.3 shows an expression for the carrier density as functions of the position of the Fermi level and the conduction and valence band. Because the carriers increase exponentially with respect to the energy level, we can write the carrier density conveniently with respect to the Fermi level and the intrinsic Fermi level (the middle of the band gap). The form of the equations is the same, but the prefactor (n_i, and N_c/N_v) and the reference value differ,

$$\begin{aligned} n_0 &= n_i \exp((E_{\text{Fermi}} - E_i)/kT) \\ p_0 &= n_i \exp((E_i - E_{\text{Fermi}})/kT) \end{aligned} \tag{6.3}$$

There is an easy way to recall Eqs. 6.2 and 6.3. Equation 6.2 says that if the Fermi level was at the conduction band (with $E_{\text{Fermi}} - E_c = 0$), then the carrier density would be N_c. Equation 6.3 references the carrier density to the intrinsic Fermi level, E_i. If the Fermi level was at the intrinsic Fermi level (with $E_{\text{Fermi}} - E_i = 0$), then the carrier density would be n_i.

A visual representation of the Fermi level, and these formulas, is shown in Fig. 6.1.

Some material constants to be used in the examples, and in the end-of-chapter problems, are tabulated in Table 6.2.

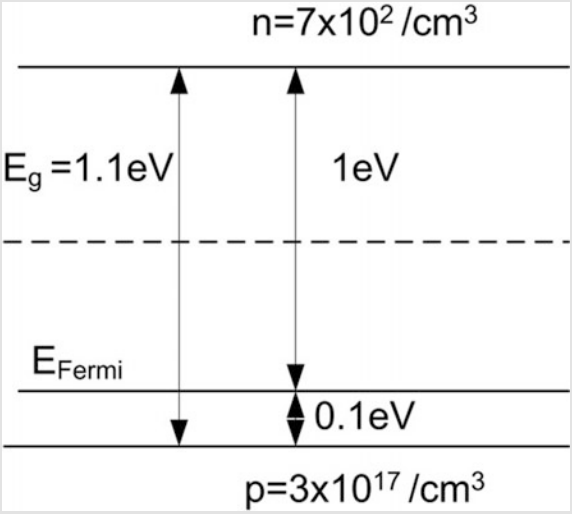
An example will illustrate the use of these equations.

Example: A Si wafer is doped with 3×10^{17} atoms/cm^3 of B. Sketch the band structure, indicating the distance between the Fermi level and the intrinsic Fermi level, and the distance between the Fermi level and the valence and conduction band. Find n_0 and p_0.

Solution: Using Eq. 6.3, and assuming $p_0 = 3 \times 10^{17}/\text{cm}^3$, then $(E_{\text{Fermi}} - E_i) = kT \ln(N_A/n_i) = 0.026 \ln(3 \times 10^{17}/10^{10}) = 0.45$ eV from the intrinsic Fermi level. The band gap of Si is 1.1 eV, so if the Fermi level is 0.45 eV from

the middle (0.55 eV), then it is about 0.1 eV from the valence band and 1 eV from the conduction band.

Just to illustrate, N_v for Si is $1 \times 10^{19}/cm^3$. From Eq. 6.1, $3 \times 10^{17} = 1 \times 10^{19} \exp(-(E_v - E_F)/0.026))$, or $E_v - E_F = 0.09$ eV, which is approximately the same value. The numbers, n_0 and p_0, can be found from Eq. 6.1 or Eq. 6.3, but most conveniently from Eq. 6.2. The term p_0 at room temperature is the doping density, $3 \times 10^{17}/cm^3$, so $n_0 = n_i^2/p_0 = (1.45 \times 10^{10})^2/3 \times 10^{17} = 700/cm^3$.



Let us also define two more useful terms. In a doped semiconductor, the majority carriers are those directly derived from the dopants (electrons from a donor-doped semiconductor), and the minority carriers are the other species, whose concentration is reduced. In the previous example, holes are the majority carriers, and electrons are the minority carriers.

6.2.2 Band Structure and Charges in *p–n* Junction

Having introduced a single semiconductor in Fig. 6.1, let us look at the properties of something more complicated. In Fig. 6.2, we show a *p–n* junction, drawn in equilibrium, as the basis for the discussion for the next several sections.

In equilibrium, there is only one Fermi level which describes the entire structure, shown stretching across from one side to another. The distance between the Fermi

level and the valence and conduction band, respectively, give the number of mobile electrons or holes in the band. Also shown in the figure is the resulting fixed charge at the junction, the direction of the electric field (and corresponding drift current), and the electric field.

Far away from the junction between the *n*- and *p*-region, the semiconductors look like *n*-doped or *p*-doped semiconductors. Here, Eqs. 6.1–6.3 apply. For example, on the *n*-side, the electron density is about equal to the dopant density, the hole density is n_i^2/N_D, and the Fermi level is near the conduction band. What happens at the junction is discussed next.

These regions on the *n*- and *p*-side are called the quasi-neutral regions. They are electrically neutral because the large number of mobile electrons comes from dopant atoms. Each mobile electron with a negative charge leaves behind a fixed positive charge dopant atom. Hence, the net charge is zero, and it is electrically neutral.

The region in the middle, where the Fermi level is far from both the conduction and valence band, has few mobile carriers but still has the immobile charge associated with the dopant atoms. This is called the space charge region or the depletion region.

Where did the mobile charges go? At the junction between the electron-rich *n*-doped side and the hole-rich *p*-doped side, the free electrons and holes recombined and vanished, leaving the space charge behind.

At the junction of these two regions, there is a very short region in which the semiconductor goes from being quasi-neutral, with zero net charge, to having many fewer mobile carriers and an electric field. This length is of the order of the Debye length, L_D, given by

$$L_D = \sqrt{\frac{\varepsilon kT}{Nq^2}} \tag{6.4}$$

where N is the dopant density, ε is the dielectric constant, and q is the fundamental charge unit.

Even for relatively low dopant densities, the Debye length is quite small. The usual assumption is of an abrupt junction between the quasi-neutral region and the depletion region, which is quite reasonable.

We can now look at the band structure of Fig. 6.2 and sketch the free charge density.

Example: Using the distance between the Fermi level and the band in Fig. 6.2, sketch the mobile charge concentration.

Solution: Far away from the junction, the free carrier concentration of electrons and holes is equal to the dopant density. In the depletion region, the Fermi level

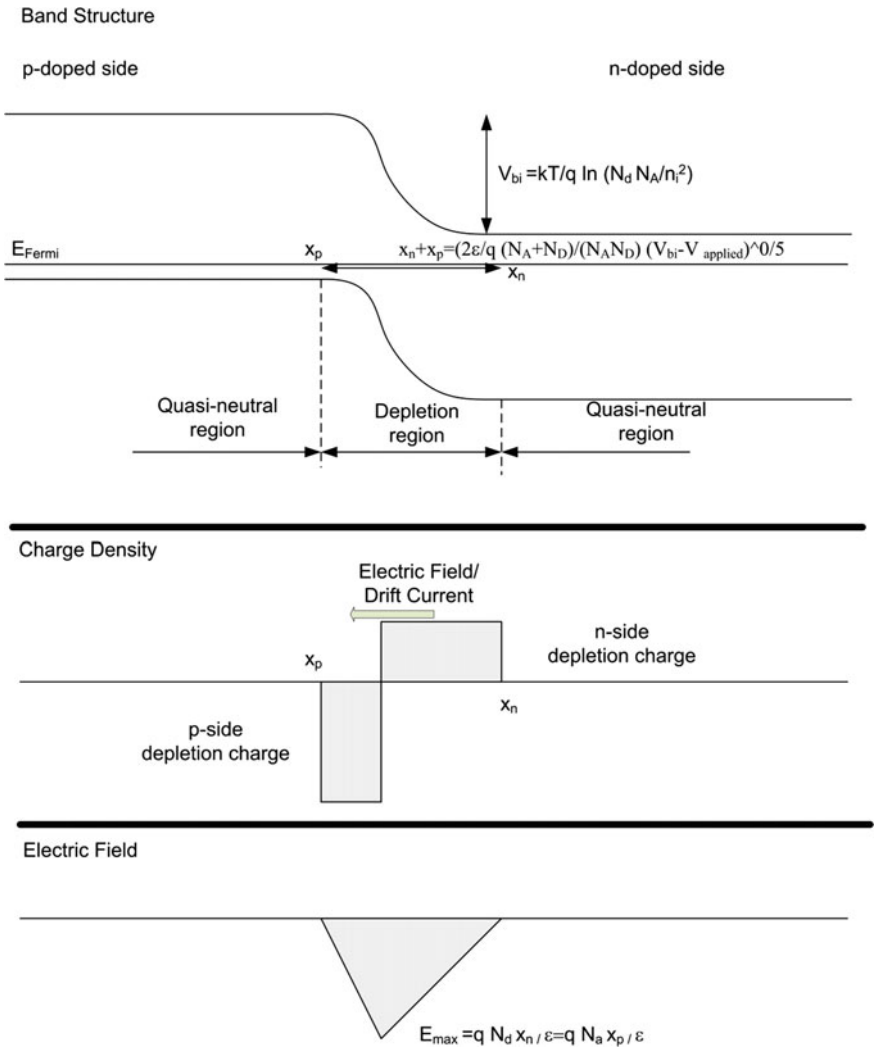
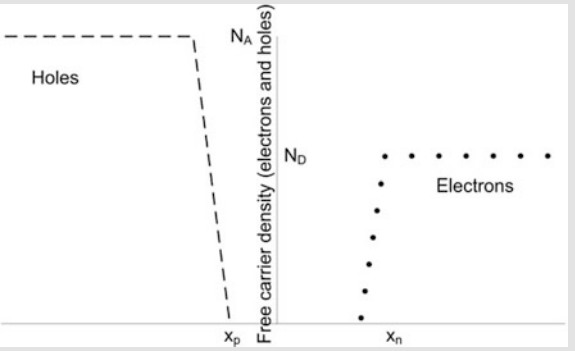


Fig. 6.2 Band structure, depletion charge density, and electric field of a *p–n* junction in equilibrium. Some equations to be developed are already shown in the diagram

is far from both the conduction and valence bands, leading to a very low concentration of both electrons and holes. The holes and electrons, brought in close proximity, recombine. The overall sketch of free carrier density is given below.



To summarize, there are:

(i) Mostly mobile electrons on the *n*-side of the junction balanced by ionized dopants;
(ii) Mostly mobile holes on the *p*-side of the junction balanced by the ionized dopants; and
(iii) Very few mobile electrons or holes in the middle of the junction (the space charge region).

Because the space charge region is charged, it has an electric field associated with it. The electric field always points from positive charge to negative charge. In this case, it points from the *n*-side (which has positive space charge) to the *p*-side (which has negative space charge).

6.2.3 Currents in an Unbiased *p–n* Junction

6.2.3.1 Diffusion Current

In a *p–n* junction under no applied voltage, there is no net current, However, there are current components. In particular, on the one side of the junction (the *n*-side), there are a lot more electrons than there are on the other side (the *p*-side). There is a diffusion of electrons from the electron-rich *n*-side to the *p*-side. Diffusion current in general is given by

$$J_{p-\text{diffusion}} = qD_p\left(\frac{\mathrm{d}p}{\mathrm{d}x}\right)$$
$$J_{n-\text{diffusion}} = -qD_n\left(\frac{\mathrm{d}n}{\mathrm{d}x}\right) \tag{6.5}$$

where J is the diffusion current, n and p are the concentrations of electrons or holes, respectively, and q is the fundamental unit of charge. The current is proportional to the difference in carrier concentration ($\mathrm{d}n/\mathrm{d}x$) with a proportionality constant D that depends on the material and on the carrier (holes or electrons). The change in sign between electrons and holes is simply related to the charge of the carrier.

This expression makes common sense; if you put a drop of cream into coffee, the entire cup of coffee gradually gets lighter as the cream diffuses from regions where there is more cream (where it is first dropped in) to regions where there is less cream. Random motion provided by temperature serves to spread out things from regions of high concentration to low concentration.

In a *p–n* junction, we expect there to be some diffusion current associated with holes moving from the *p*-side to the *n*-side (current going to the right) and with electrons moving from the right to the left (also positive current going to the right).

6.2.3.2 Drift Current

There is also a built-in electric field associated with the space charge region. The electric field points from the *n*-side to the *p*-side. That means that any mobile charge carriers that happen to fall into the space charge region will be caught by that electric field and swept to one side or another. The formula for drift current is

$$J_{n-\text{drift}} = -qE\mu_n n$$
$$J_{p-\text{drift}} = qE\mu_p p \tag{6.6}$$

where E is the electric field, and μ is the mobility of electrons or holes, respectively. The reader is reminded that the mobility μ is related to the diffusion current, D, through the Einstein relation

$$\frac{D}{\mu} = \frac{kT}{q} \tag{6.7}$$

Fundamentally, the reason is that electrical mobility, and diffusion, both involve carriers scattering randomly off of atoms in a crystal lattice. With an electric field, there is displacement due to the electric field between collisions, which essentially resets the direction of travel of the carrier; with diffusion, the random motion is always random, but adds up to movement of the carriers from regions of high concentration to low concentration. This will be explored further in the problems.

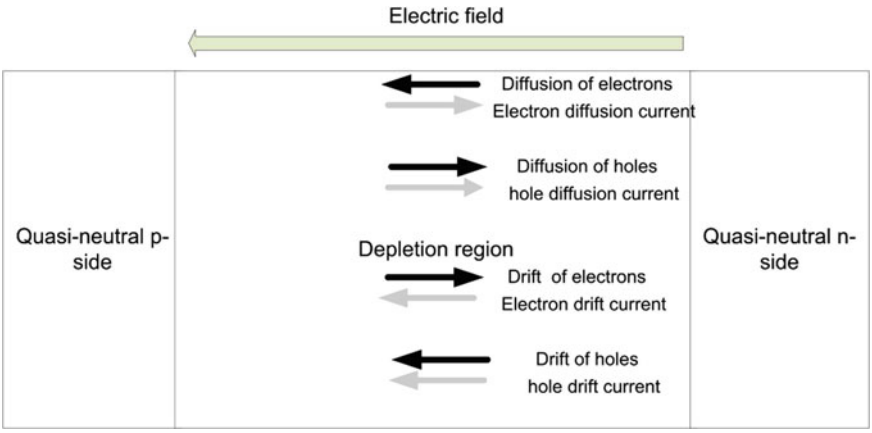


Fig. 6.3 Current components across a *p–n* junction in equilibrium

The drift direction in which the carriers will go is interesting. From the *n*-side of the quasi-neutral region, minority carriers (holes) which happen to fall into the space charge region will drift over toward the *p*-side; similarly, minority electrons on the *p*-side will drift over to the *n*-side. The drift current is in the opposite direction to the diffusion current. At equilibrium, the net current is zero.

The drift and diffusion currents in a *p–n* junction in equilibrium are shown in Fig. 6.3.

Questions about *p–n* junctions are very common on qualifier examinations for Ph.D. students. As an aid for working out directions, the author suggests considering diffusion first. Diffusion is more intuitive (electrons of course diffuse from the region with high electron concentration, the *n*-side, to the *p*-side), and drift current is in the other direction. Remember to change the sign of the current direction when the moving charge is negative!

6.2.4 Built-in Voltage

Figure 6.2 shows that an electron or hole is at a different energy level on one side of the junction than the other. This difference is called the built-in voltage and is determined by the difference in the doping levels on each side of the device.

A simple expression for the built-in voltage can be worked out from Eq. 6.2. The carrier density on each side of the junction is approximately equal to the dopant density at room temperature,

$$\begin{aligned} N_{\mathrm{d}} &= n_i \exp((E_{\mathrm{Fermi}} - E_i)/kT) \\ N_{\mathrm{a}} &= n_i \exp((E_i - E_{\mathrm{Fermi}})/kT) \end{aligned} \tag{6.8}$$

where N_d and N_a are the dopant densities of donors (n-side) and acceptors (p-side), respectively. These expressions can be rearranged to be

$$\begin{aligned} E_f - E_i &= kT \ln\left(\frac{N_d}{n_i}\right) \\ E_i - E_f &= kT \ln\left(\frac{N_a}{n_i}\right) \end{aligned} \tag{6.9}$$

The first expression tells how much the conduction band is above the Fermi level on the n-side. The second expression tells how much the valence band is below the Fermi level on the p-side. From Fig. 6.2, it should be clear that the sum of these two expressions (given that the Fermi level is a fixed reference) is the built-in voltage, V_{bi},

$$V_{bi} = \frac{kT}{q} \ln\left(\frac{N_d N_a}{n_i^2}\right) \tag{6.10}$$

6.2.5 Width of Space Charge Region

The built-in voltage above is created by the space charge left in the space charge region. Since we know the built-in voltage, and the charge density, we can determine the width of this space charge region, as described below.

The relationships between charge density, ρ, and electric field, E, are

$$\frac{dE}{dx} = \frac{\rho}{\varepsilon} = \frac{qN_{A/D}}{\varepsilon} \tag{6.11}$$

$$E = \int_{x_p}^{x_n} \frac{\rho(x)}{\varepsilon} dx \tag{6.12}$$

where ε is the dielectric constant, and N is the dopant density of acceptors or donors (with the sign of the charge appropriately matching). This electric field is illustrated in Fig. 6.2. For an abrupt junction with a constant charge density on each side, the electric field is a maximum at the junction and falls to zero outside the depletion region.

The electric field is the integral of the space charge density. In general, it is easiest to keep the signs straight by just recalling that electric field points from positive to negative charges. The integral goes from the (currently unknown) left edge of the space charge region, x_n, where it starts at zero, to the right edge of the

space charge region, where it ends at zero again at x_p. The electric field is maximum right at the junction between the p- and the n-sides. The maximum electric field $E_{\max}$ is

$$E_{\max} = \frac{qN_A x_p}{\varepsilon} = \frac{qN_D x_n}{\varepsilon} \tag{6.13}$$

With the electric field determined, the voltage is simply the integral of the electrical field.

$$V_{bi} = \int_{x_p}^{x_n} E(x)\mathrm{d}x \tag{6.14}$$

There is one other relationship between x_p and x_n that we can use. The total amount of depletion charge has to be zero (why?). This relationship can be expressed as

$$x_p N_A = x_n N_D \tag{6.15}$$

Using Eqs. 6.13–6.15, the depletion layer width can be expressed in terms of the doping as

$$x_p + x_n = \sqrt{\frac{2\varepsilon}{q}\frac{N_A + N_D}{N_A N_D}\left(V_{bi} - V_{applied}\right)} \tag{6.16}$$

where V_{bi} is the built-in voltage, and $V_{applied}$ is the applied bias (which we will talk about in the next section).

For a junction with an abrupt change between p-dopants and n-dopants, this is the appropriate formula. For other dopant formulations (for example, a linear gradient making a smooth transition from a p-side to an n-side), different formulas can be derived, all of them based on the idea of a built-in voltage between one side and the other, and a region completely depleted of mobile charges sandwiched between quasi-neutral regions that are charge neutral.

A few qualitative observations are helpful. First, Eqs. 6.15 and 6.16 describe how much of the depletion layer width appears on each side of the junction. Because of overall charge neutrality, the width of the depletion layer is wider on the more lightly doped side of the junction. If $N_D = 10N_A$, for example, the depletion layer width will be 10 times larger on the p-side than on the n-doped side. If one doping is significantly greater than the other (say 10× or more), it is usually

accurate enough to assume that all the depletion width appears on the lightly doped side.

Another qualitative observation is that in a laser with an undoped active region (or a *p–i–n*) diode, the middle section is undoped. The undoped middle section looks like part of the depletion region in the sense of having relatively few mobile charges. Depleted *n*- and *p*-layers appear at the edges of the doped active regions, but the bulk of the built-in voltage is taken up by the voltage drop across the undoped region. We will explore this further in the problems. Meanwhile, let us do an example of the application of these equations.

Example: A Si abrupt junction is formed between a *p*-doped $10^{18}/\text{cm}^3$ region and an *n*-doped $5 \times 10^{16}/\text{cm}^3$ region. Sketch the band structure, labeling the distance between the Fermi level and the conduction and valence band on each side. Find the width of the depletion region on both the *n*- and the *p*-side. Find the built-in voltage and the peak electric field and indicate its direction.

Solution: Start by drawing a straight line indicating the Fermi level in equilibrium. From Eq. 6.8, the Fermi level is $E_f - E_i = kT \ln\left(\frac{5\times10^{16}}{1.45\times10^{10}}\right) = 0.37\,\text{eV}$ above the intrinsic Fermi level on the *n*-side and $kT \ln\left(\frac{10^{18}}{1.45\times10^{10}}\right) = 0.47\,\text{eV}$ below the intrinsic Fermi level on the *p*-side. The built-in voltage is then $V_{\text{bi}} = 0.37\ \text{eV} + 0.47\ \text{eV} = 0.84\ \text{eV}$.
The width of the depletion region is then Eq. 6.16,

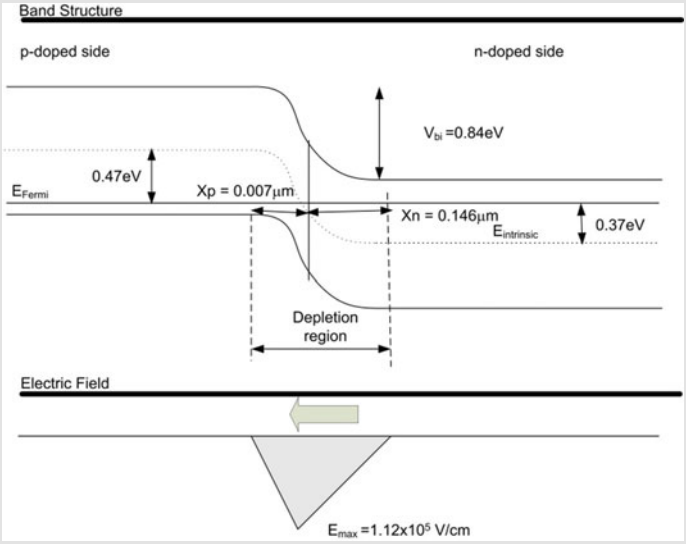
$$x_n + x_p = \sqrt{\frac{2(11.7)(8.854\times10^{-14})}{1.6\times10^{-19}}\frac{(5\times10^{16}+10^{18})}{(5\times10^{16})(10^{18})}(0.84)}$$
$$= 0.153\,\mu\text{m}$$

Now, because the *n*-doping density is 20× less than the *p*-doping density, practically all of this is on the *n*-side. However, to work it out properly, we have two equations: $5 \times 10^{16} x_n = 10^{18} x_p$, and $x_p + x_n = 0.153\ \mu\text{m}$, gives $x_p = 0.007\ \mu\text{m}$ and $x_n = 0.146\ \mu\text{m}$.

The peak electric field is given by Eq. 6.13 and is $1.6 \times 10^{-19}(5 \times 10^{16}\ /\text{cm}^3)(0.146 \times 10^{-4}\text{cm})/(8.854 \times 10^{-14}\,\text{F/cm})(11.7) = 1.12 \times 10^5\ \text{V/cm}$. It points from *n*-side to *p*-side.

The only care to be taken is with the units. Since constants such as ε are used here, be sure to use the

constants associated with the units (for example, $\varepsilon_0 = 8.854 \times 10^{-14}$ F/cm).
Putting all this information in a diagram like Fig. 6.1 gives



6.3 Semiconductor *p–n* Junctions with Applied Bias

6.3.1 Applied Bias and Quasi-Fermi Levels

Let us now examine the diode under an applied bias V_{applied} (where a voltage is applied to the *p*-side, and the *n*-side is grounded). The band diagram for this diode under bias is shown below. Since it is forward biased, the barrier height shrinks, and a positive current flows from the *p*-side to the *n*-side. Since the barrier height ($V_{\text{bi}} - V_{\text{applied}}$) is lowered, the depletion layer width is reduced as well.

When this bias is applied to the *p*-side, current starts to flow. Since it is the diffusion current which flows from the *p*-side to the *n*-side, it must be the diffusion current which increases as the voltage increases. In fact, this does make sense. Drift current is composed of minority carriers which happen to wander into the depletion region and are swept to the majority carrier side. Regardless of the size of the depletion region, about the same number of minority carriers find themselves caught in the depletion region and become drift current.

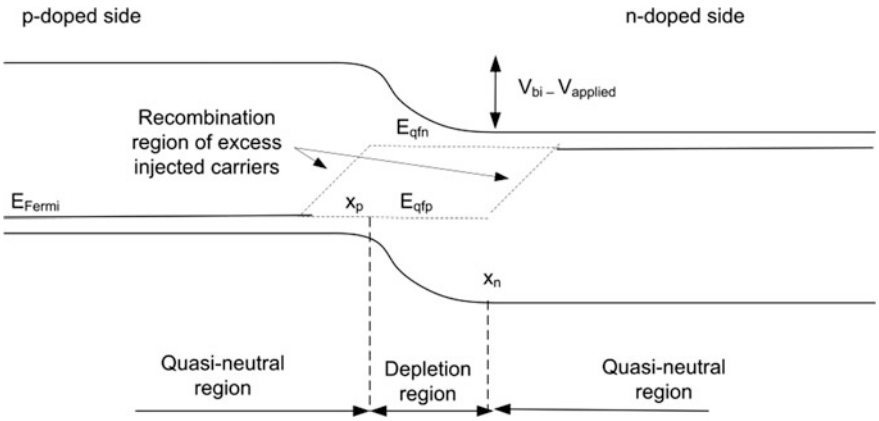


Fig. 6.4 Forward-biased *p–n* junction. The quasi-Fermi level splits, with excess electrons injected across the junction from the *n*-side and excess holes injected across the junction from the *p*-side, in the other direction

In the band diagram of Fig. 6.4, the best representation of the device under bias is with quasi-Fermi levels. (As we talked about in Chap. 4, quasi-Fermi levels are separate Fermi levels for holes and electrons.) Far from the junction on the right side, the semiconductor is by itself in equilibrium. Because there is a bias applied, more holes are injected into the depletion region. Assuming minimal recombination as they make their way across, these excess carriers appear at the edge of the *p*-side quasi-neutral region. In the quasi-neutral region, these excess minority carrier holes recombine with the majority carrier electrons until equilibrium is restored on the left side. Again, far from the junction on the left side, the semiconductor is back in equilibrium, with only one Fermi level.

The best way to draw the band structure is to draw both the left side and the right side with the Fermi levels located as appropriate, and then separate them by the applied voltage V_{applied}. Then label the *p*-side Fermi level $E_{\text{qf}p}$ and extend it into the *n*-side; label the *n*-side Fermi level $E_{\text{qf}n}$ and extend it into the *p*-side. At the boundary of the *n*-side depletion region, the carriers enter a region with high carrier density again and start recombining as they diffuse. As the minority carriers on each side diminish, the quasi-Fermi levels approach each other again.

Looking at the quasi-Fermi levels, we can sketch the free carrier density in the quasi-neutral region.

Far away from the junction, the carrier density is the intrinsic carrier density with that doping density. Near the border of the depletion region, the quasi-Fermi levels split, and there starts to be an excess of minority carriers. (There is also the same number of excess majority carriers to maintain quasi-neutrality. However, the percentage change in minority carrier density is much, much greater.)

Across the depletion region, there are more electrons and holes than there would be in equilibrium. However, it is assumed that the carrier density is still too low for significant recombination, so the extra carriers on each side are injected across the depletion region and appear on the other side.

6.3.2 Recombination and Boundary Conditions

Let us go from the band structure in Fig. 6.4 and charge density in Fig. 6.5 to the current density. We know there is no current with no applied bias, and we wish to determine the current with an applied bias. For reasons that will hopefully become clear in the next section or two, let us focus on the diffusion of minority carriers in the quasi-neutral region.

Given the band structure of Fig. 6.4, and the carrier density of Fig. 6.5, the density of minority carriers at the edge of the quasi-neutral region is given as

$$\begin{aligned} n_p &= n_{p0} \exp\left(qV_{\text{applied}}/kT\right) \\ p_n &= p_{n0} \exp\left(qV_{\text{applied}}/kT\right) \end{aligned} \tag{6.17}$$

where n_p and p_n are the minority carrier density at the edge of the quasi-neutral region, and n_{p0} and p_{n0} are the minority carriers in equilibrium with the same doping density. The carrier density, of course, depends exponentially on the Fermi levels. The equilibrium densities of minority carriers, n on the p-side (n_{p0}) and p on the n-side (p_{n0}), are given by

$$\begin{aligned} n_{p0} &= \frac{n_i^2}{N_{\text{A}}} \\ p_{n0} &= \frac{n_i^2}{N_{\text{D}}} \end{aligned} \tag{6.18}$$

which is Eq. 6.2, with n or p equal to N_{D} or N_{A}.

Look closely at the n-side, where the minority carriers are holes. At the edge, there is an excess number of holes; far from the boundary, everything has returned to equilibrium. Therefore, there is a diffusion of minority holes into the n-side. As

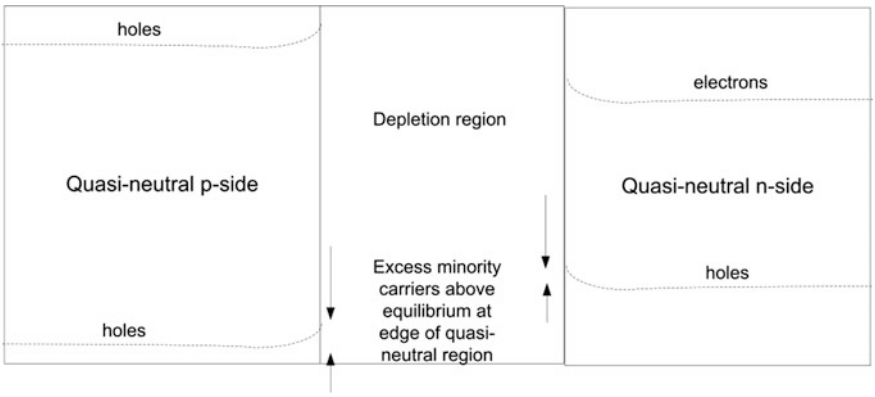


Fig. 6.5 Mobile charge density of holes and electrons in the quasi-neutral region under forward bias. Note that there are more electrons and holes on both sides of the depletion region

these excess minority (and majority) carriers diffuse away from the junction, they recombine, until they return to equilibrium. There are still minority carriers, but they are now in thermal equilibrium with the majority carriers. The amount of minority carriers generated thermally is equal to the amount disappearing through recombination.

The equations for excess minority carriers can be most conveniently written by defining a variable Δn, which is the number of minority carriers above equilibrium,

$$\begin{aligned} \Delta n_p &= n_{p0}\left(\exp\left(qV_{\text{applied}}/kT\right) - 1\right) \\ \Delta p_n &= p_{n0}\left(\exp\left(qV_{\text{applied}}/kT\right) - 1\right) \end{aligned} \tag{6.19}$$

The equation below describes the combined diffusion and recombination of carriers in the active region. We are interested in the steady state solution when the concentrations are not changing with time,

$$\frac{\mathrm{d}\Delta n(x,t)}{\mathrm{d}t} = 0 = D\frac{\mathrm{d}^2\Delta n(x,t)}{\mathrm{d}x^2} - \frac{\Delta n(x,t)}{\tau} \tag{6.20}$$

This comes from Fick's second law of diffusion and conservation of particles. In this expression, D is the diffusion coefficient, and τ is the carrier recombination lifetime. In other words, what it says is that the change in concentration for any given point $n(x)$ depends on the flux of carriers in, the flux of carriers out, and recombination.

There can also be a current component due to generation (in semiconductors, if the number of carriers is below the equilibrium number, carriers are thermally generated in the material. We neglect it in this equation). The equation is shown pictorially below. The excess holes both recombine, and diffuse, in the quasi-neutral region.

Taking the coordinates as sketched in Fig. 6.6, the boundary conditions for this differential equation are

$$\Delta p_n(0) = p_{n0}\left(\exp\left(qV_{\text{applied}}/kT\right) - 1\right) \tag{6.21}$$

and

$$\Delta p_n(\infty) = 0 \tag{6.22}$$

(The minority concentration returns to equilibrium far from the junction). With these equations and boundary conditions, the solution $\Delta p_n(x)$ is

$$\Delta p_n = p_{n0}\exp\left(-x/\sqrt{D\tau}\right)\left(\exp\left(qV_{\text{applied}}/kT\right) - 1\right) \tag{6.23}$$

The term $\sqrt{D\tau}$ appears in this equation. This term has dimensions of length and is called the diffusion length, L_{D}. It represents the typical length that a carrier will travel before it recombines. Equation 6.24 gives the diffusion length for electrons

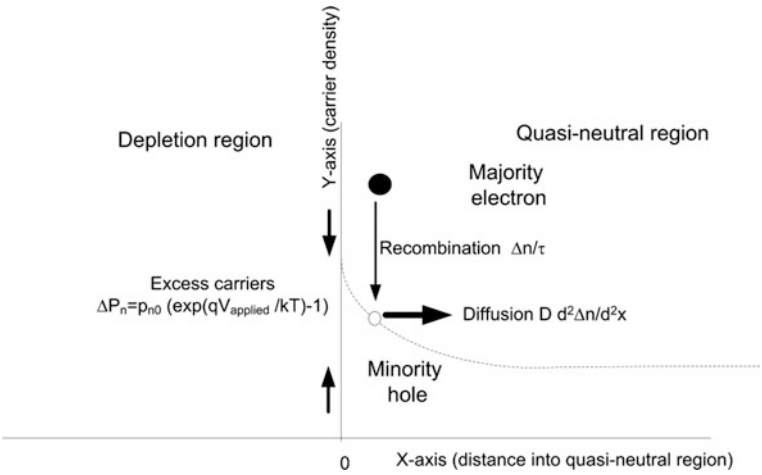


Fig. 6.6 Diffusion current at the edge of the quasi-neutral region, showing the holes diffusing and recombining as they diffuse away from the junction

and holes, written with subscripts as a reminder to use the appropriate lifetime and diffusion coefficient for each carrier on the correct side of the junction.

$$\begin{aligned} L_n &= \sqrt{D_n \tau_n} \\ L_p &= \sqrt{D_p \tau_p} \end{aligned} \tag{6.24}$$

6.3.3 Minority Carrier Quasi-Neutral Region Diffusion Current

Finally, from Eq. 6.5, we are in a position to calculate the current: specifically, the diffusion current associated with minority carriers on the *n*-side of the junction.

Equation 6.23 gives the excess carrier concentration, $\Delta p_n(x)$. From Fick's law, the diffusion current of minority carriers on the *n*-side is proportional to

$$J = qD\frac{\mathrm{d}\Delta p_n}{\mathrm{d}x} = qD\frac{p_{n0}}{\sqrt{D\tau}}\exp\left(-x/\sqrt{Dt}\right)\left(\exp\left(qV_{\text{applied}}/kT\right) - 1\right) \tag{6.25}$$

where x, we remind the reader, is the distance from the edge of the depletion region going into the quasi-neutral region. An identical equation can be derived for electron minority current on the *p*-side. The current density J here is the current density in A/cm^2 in cross-sectional area.

Now, finally, we are in a position to write down the diode current equation. Before we do, to make it realistic, we have to add a few more subscripts. The diffusion coefficient is different for electrons and holes (for one thing, the mobility

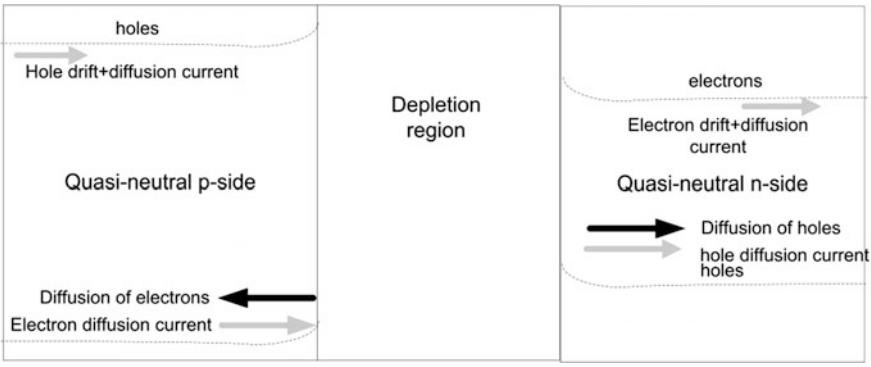


Fig. 6.7 Current components in the quasi-neutral regions of a forward-biased diode

for electrons is different from the mobility for holes, and according to the Einstein relation, that means the diffusion coefficient will be different as well). In fact, the diffusion coefficient depends not only on whether it is holes or electrons which are diffusing, but also on the ambient dopant density, which depends on which side of the junction the diffusion takes place. We will label the diffusions, $D_{n\text{–}p\text{-side}}$ and $D_{p\text{–}n\text{-side}}$ to refer to the diffusion of (minority carrier) electrons on the *p*-side or diffusion of (minority carrier) holes on *n*-side.

The lifetime of electrons or holes is also different, so we will now label τ as τ_p and τ_n.

Now, let us think about currents in a more qualitative way, as illustrated in Fig. 6.7. Current has to be continuous across the device, since there is no charge accumulation. We know what charge distribution looks like across the device under an applied bias, that is, given from Fig. 6.5. Based on the derivative of charge distribution, we can label currents in the charge picture shown in Fig. 6.7.

Across and up to the edges of the depletion region, there is no meaningful recombination; therefore, both electron and hole currents have to be separately continuous. The majority carrier current on each side is actually carried by a combination of drift and diffusion (once the charge distribution has reached equilibrium, there can be no more diffusion current; drift is much more significant for majority carriers because the current is proportional to the number of carriers).

On the left side of the junction, the electron current is all diffusion of minority carriers. On the right side of the junction, all the hole current is diffusion current of minority carriers. Therefore, the total current across the junction is the minority carrier current at the edge of the *n*-side plus the minority carrier diffusion current at the edge of the *p*-side. Written down, it is

$$J = q\left(D_{p-n\,\text{side}}\frac{p_{n0}}{\sqrt{D\tau_p}} + D_{n-p\,\text{side}}\frac{p_{p0}}{\sqrt{D\tau_n}}\right)\left(\exp\left(qV_{\text{applied}}/kT\right) - 1\right) \qquad (6.26)$$

Written to put it in terms of the intrinsic number of carriers in the semiconductor (n_i) and the doping level, the equation can be written as

$$J = q\left(D_{p-n\,\text{side}}\frac{n_i^2}{N_\text{D}\sqrt{D\tau_p}} + D_{n-p\,\text{side}}\frac{n_i^2}{N_\text{A}\sqrt{D\tau_n}}\right)\left(\exp\left(qV_\text{applied}/kT\right) - 1\right) \quad (6.27)$$

or it is sometimes written as

$$J = q\left(D_{p-n\,\text{side}}\frac{n_i^2}{N_\text{D}L_{p-n\,\text{side}}} + D_{n-p\,\text{side}}\frac{n_i^2}{N_\text{A}L_{n-p\,\text{side}}}\right)\left(\exp\left(qV_\text{applied}/kT\right) - 1\right). \quad (6.28)$$

However, most people will recognize it most easily as the diode equation,

$$\begin{aligned} J_0 &= q\left(D_{p-n\,\text{side}}\frac{n_i^2}{N_D L_{p-n\,\text{side}}} + D_{n-p\,\text{side}}\frac{n_i^2}{N_A L_{n-p\,\text{side}}}\right) \\ J &= J_0(\exp(qV_\text{applied}/kT) - 1) \end{aligned} \quad (6.29)$$

in which the diode current depends exponentially on the applied voltage and a prefactor term J_0 which depends on the doping and material characteristics.

Let us now work through an example.

Example: A silicon *p–n* junction has the following characteristics.

n-side	*p*-side
μ_n = 1000 cm^2/V s μ_p = 400 cm^2/V s	μ_n = 500 cm^2/V s μ_p = 180 cm^2/V s
τ_n = 500 μS τ_p = 30 μS	τ_n = 10 μs τ_p = 1 μs
$N_\text{D} = 5 \times 10^{16}$/cm^3	$N_\text{A} = 10^{18}$/cm^3

Find the diffusion lengths, L_p and L_n, and the reverse saturation current density, J_0.

Solution: This is Eq. 6.16, where the only hard part is picking out the right constants. On the *n*-side, we are looking at the diffusion of minority holes, so the correct numbers are τ_p and D_p. D_p can be calculated from μ_p as $D_p = kT/q\,\mu_p = 0.026 \times 400 = 10.4\,\text{cm}^2/\text{s}$. On the *p*-side, similarly, the relevant numbers are τ_n and D_n, which are 10 μs and 13 cm^2/s.

The diffusion lengths then are $\sqrt{10 \times 10^{-6} * 13} = 114\,\mu\text{m}$ for electrons on the *p*-side, and $\sqrt{30 \times 10^{-6} * 10} = 176\,\mu\text{m}$ for holes on the *n*-side.
The prefactor J_0 is given by Eq. 6.29, or,

$$1.6 \times 10^{-19}\left(10\frac{(1.45 \times 10^{10})^2}{(5 \times 10^{16})(0.0176)} + 13\frac{(1.45 \times 10^{10})^2}{(10^{18})(0.0114)}\right) = 4.32 \times 10^{-12}\,\text{A/cm}^2.$$

6.4 Semiconductor Laser *p–n* Junctions

6.4.1 Diode Ideality Factor

Having reminded the reader of the *I–V* curve of an ideal abrupt *p–n* junction, let us talk about the *I–V* curve of a working laser or a real diode. There are several differences.

The ideal diode equation (Eq. 6.29) was derived neglecting currents that come from recombination, or generation, within the depletion region. Actual diodes have equations that look like Eq. 6.29, but with a diode ideality factor, n, as

$$J = J_0\left(\exp\left(qV_{\text{applied}}/nkT\right) - 1\right) \tag{6.30}$$

This ideality factor is determined by measuring the *I–V* curve of the laser and fitting it to the form of Eq. 6.30. They reflect the influence of these non-ideal terms, like recombination or generation currents. In general, most diodes have a diode ideality factor greater than 1. Laser diodes, in particular, are designed to facilitate recombination, and the ideality factor of lasers is closer to 2.

Second, a laser typically does not have an abrupt junction. Often the laser has an undoped active region, which means it has several hundreds of nanometers, or more, of undoped material. The diode looks more like a *p–i–n* junction than a *p–n* junction. That makes the peak electric field across the junction less and the effective depletion width somewhat more. (This will be explored further in the problems.)

6.4.2 Clamping of Quasi-Fermi Levels at Threshold

Above threshold, the differences are more interesting. First, let us define the differential resistance, R_{diff}, of a diode (or any device).

$$R_{\text{diff}} = \frac{\mathrm{d}V}{\mathrm{d}I} = \frac{1}{\frac{\mathrm{d}I}{\mathrm{d}V}} = \frac{kT}{I(V)} \tag{6.31}$$

This differential resistance is the reciprocal of the slope at each point. In a conventional diode, the differential resistance continually decreases.

However, the physical phenomenon on which this is based is the continual splitting of the quasi-Fermi levels as the voltage increases. In a laser, the quasi-Fermi levels are clamped above threshold; above threshold, all the extra carriers that are injected into the active region leave as photons. Because the quasi-Fermi levels are clamped, the differential resistance is also clamped. This differential resistance is actually no longer a 'diode' resistance; it represents the parasitic resistance due to the contact resistance of the metals and the ohmic resistance across the *p*- and *n*-side of the active region.

There is a fairly dramatic difference between the differential resistance curve of a conventional diode and a laser diode. Figure 6.8 shows the *I*–*V*, and differential resistance, measured from a laser, at threshold and above, along with the *I*–*V* and *I*–d*V*/d*I* curve of a fictitious diode with the same *n* and saturation current.

At threshold, the resistance of a laser drops and is constant, with a value equal to the parasitic resistance. This parasitic resistance is often a laser parameter with a product specification to be less than 10 Ω or so; the higher this value becomes, the more heat gets injected into the active region along with the current.

The differential resistance of a diode is continually decreasing. In a sense, the laser diode is no longer a diode at threshold, but has a clamped band structure. It is

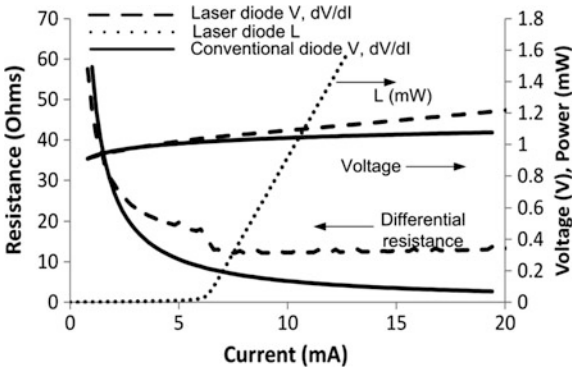


Fig. 6.8 *I*–*V*, and *I*–d*V*/d*I* curve, of a conventional diode (with matching ideality factor and reverse saturation current). The differential resistance of the conventional diode decreases with current, while the differential resistance of the laser diode is clamped

also interesting to see that the diode can be distinguished from a laser diode, and the laser diode's threshold current even measured, with a purely electrical *I–V* measurement!

6.5 Summary of Diode Characteristics

To quickly summarize Sects. 6.2–6.4, the basics of *p–n* junctions were reviewed. After the diode equation was developed, a few important differences between it and real lasers were pointed out. First, the laser quasi-Fermi levels are 'clamped' above threshold. Above threshold, the *I–V* relationship is no longer exponential, but is actually linear again. The slope (the dynamic resistance) is from the parasitic resistance due to the conduction through the semiconductor and the contact resistances from the metal contacts.

Second, the classic diode equation has a diode ideality factor $n = 1$ and neglects recombination currents in the active region. In fact, laser diodes are designed to facilitate recombination in the active region and so typically have diode ideality factors, below threshold, closer to 2.

We also note that the actual peak electric field across a laser active region is usually substantially lower than that in a *p–n* junction, because of the (generally undoped) quantum wells.

6.6 Metal Contact to Lasers

Apart from forming the *p–n* junction, the other major electrical task is to make contact with an operating laser. Since it is a semiconductor device, ultimately it has to come down to metal. The classic problem of how to get a good metal to semiconductor contact is one that was first associated with Schottky. We can start talking about the problem by drawing the band structure associated with a metal–semiconductor contact.

6.6.1 Definition of Energy Levels

Figure 6.9 shows a diagram of a metal–semiconductor contact in equilibrium. This is a Schottky junction (which we distinguish from an ohmic contact, which we will talk about in Sect. 6.7). We are going to discuss energy levels, so let us quickly define a few more levels that are relevant to the metal and to the junction.

The vacuum level is simply the energy of a free carrier which is not interacting with the material—for example, an electron above a metal surface. The energy level is labeled E_0 in the diagram. The metal work function ($q\Phi_m$) is the energy from the Fermi level in the metal to this vacuum level. This represents the amount of energy

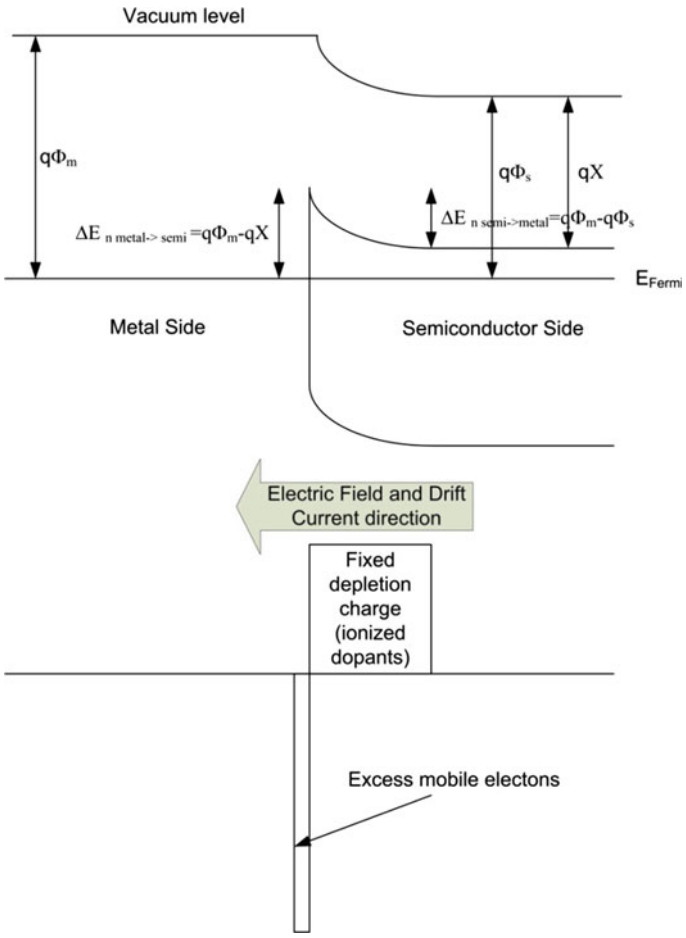


Fig. 6.9 Top, a semiconductor–metal band diagram, showing the metal work function and electron affinity. Bottom, the charge in a metal–semiconductor junction

it takes to remove one electron from the material. This is a material constant which varies for different metals.

The band structure of a metal is fairly simple. Unlike a semiconductor, a simple metal has plenty of states both below and above the Fermi level. To a good approximation, all of the states below the Fermi level are occupied, and all of the states above the Fermi level are empty.

A similar, yet different, quantity from the metal work function is the electron affinity, qX, of a semiconductor. The electron affinity is the energy distance between the conduction band and the vacuum level, and it represents the energy necessary to remove an electron from the semiconductor. This is the relevant material constant for semiconductors. The electron affinity of Si, for example, is 4.35 eV.

Semiconductors also have a work function, $q\Phi_s$, or distance from the Fermi level to the vacuum level. This is less relevant than in a metal, because typically there are no carriers at the level to be ionized. Nor is it a material constant; the distance between the semiconductor work function and the Fermi level depends on the doping. For n-doped semiconductors, it is

$$q\Phi_s = qX + kT \ln\left(\frac{N_d}{n_i}\right) \tag{6.32}$$

The junction between the metal and semiconductor is characterized by barriers. For electrons, from metal to semiconductors, the barrier height is given by,

$$\Delta E_{n\ \text{metal}\rightarrow\text{semi}} = q\phi_{ms} = q\phi_m - qX \tag{6.33}$$

This is a material constant and is labeled in Fig. 6.9. The other barrier to charge conduction is from the semiconductor to the metal and that relates to the amount of band bending: whether the conduction or valence bands need to bend up, or down, in order to make the vacuum level continuous. This bending is given by,

$$q\Phi_{sm} = q(\Phi_m - \Phi_s), \tag{6.34}$$

where a positive number means that it bends up, and a negative number means that it bends down. As illustrated in the diagram, this bending (in this case) is the potential energy barrier that majority carriers have going from a semiconductor to a metal.

6.6.2 Band Structures

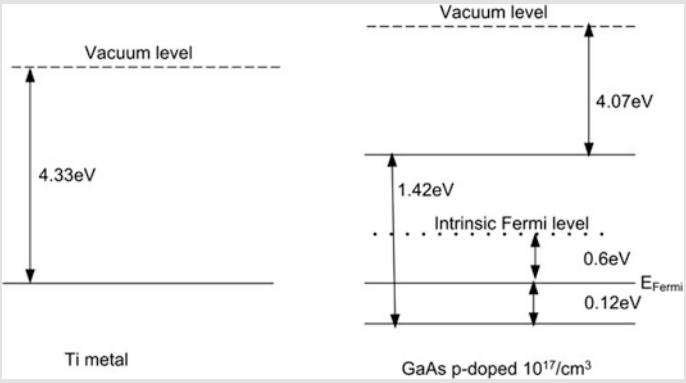
Let us discuss then how the band diagram of Fig. 6.9 is drawn and how it tells the charge distribution, both mobile and fixed.

First, the metal is specified only by the work function, $q\Phi_m$, and the semiconductor is specified by its electron affinity and the placement of its Fermi level.

To draw the band diagram when the semiconductor and metal are placed in contact, we need two guidelines. First, when they are placed in contact, everything eventually achieves equilibrium, and the band diagram starts by having a straight Fermi level across the metal and the semiconductor. A system in thermal equilibrium means that the Fermi level is constant. The second constraint is that the vacuum level is everywhere continuous. This is a physically reasonable guideline; if the vacuum level was not continuous, then a carrier could be ionized, moved a tiny little bit (from the metal side to the semiconductor side), and somehow acquire or lose energy.

Example: Sketch the band diagram of the semiconductor/metal junction given.

$$\mathrm{GaAs}\left(p = 10^{17}/\mathrm{cm}^3, X = 4.07\,\mathrm{eV}\right) \text{ to } \mathrm{Ti}(\Phi_m = 4.33\,\mathrm{eV})$$



Far away from the junction, the semiconductor and metal look like they do in free space. Following the example in Sect. 6.2.1, the location of the Fermi level is placed 0.12 eV above the valence band.

At the junction, we draw the bands assuming that the vacuum level is continuous. At the junction, the distance from the conduction band to the vacuum level is qX; the distance from the metal work function to the vacuum level is $q\Phi_m$. Therefore, the barrier for electrons from the metal to the conduction band is

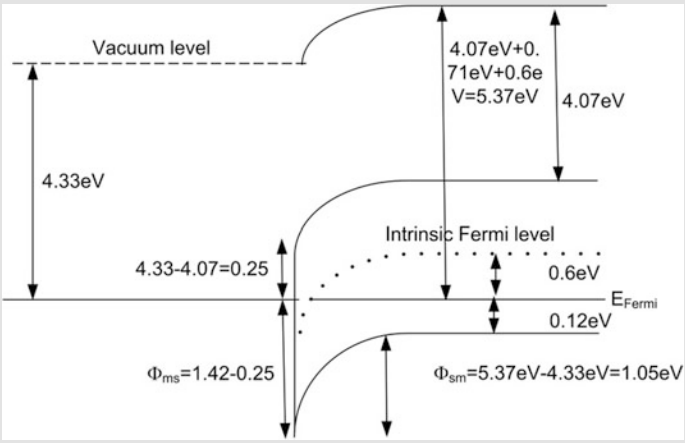
$$\Delta E_{n\,\mathrm{metal}\to\mathrm{semi}} = q\phi_m - q\mathrm{X} = 4.33 - 4.07 = 0.25,$$

which is independent of the doping and depends instead only on the metal work function and semiconductor electron affinity.

In this case, the conductors are holes; therefore, the appropriate barrier to identify is the barrier to holes (which is $E = E_g - \Delta E_{n\,\mathrm{metal}\to\mathrm{semi}}$). With this information, we can draw the junction points—line up the Fermi levels and locate the conduction and valence bands according to the barriers given.

Finally, we have to identify how much the bands bend and in what direction. The work function for the semiconductor is 5.37 eV (4.07 eV + 1.42 eV − 0.12 eV).

According to Eq. 6.34, the barrier is $q\Phi_{sm} = q(\Phi_m - \Phi_s) = 4.33 - 5.37 = -1.04\,\text{eV}$, with the negative number meaning it bends down. Combining all this information, the band structure looks like

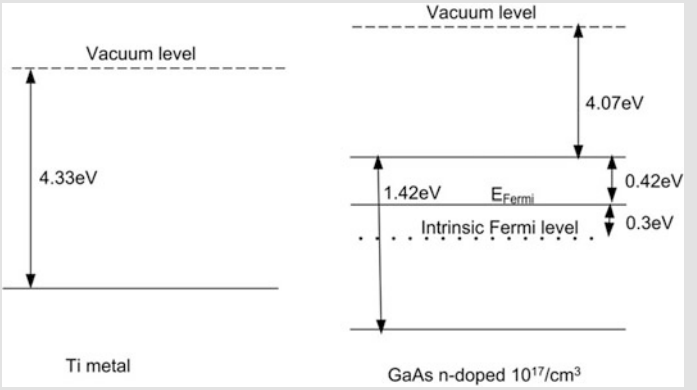


What kind of a junction is this? Well, the valence band bends away from the Fermi level in a *p*-doped material, which means a decrease in mobile carriers and a depletion region. This is also what is called a Schottky junction (a metal–semiconductor junction that looks like half of a *p*–*n* junction.) These junctions have *I*–*V* curves that look very much like diode *I*–*V* curves, with an exponential dependence of current on voltage. This is actually not the desired contact; what we would like is a metal–semiconductor contact that looks ohmic, or resistive, with a linear dependence of current on voltage.

The figure in this example is a *p*-doped Schottky junction; Fig. 6.9 shows an *n*-doped Schottky junction. Let us illustrate in the next example an ohmic contact, in which there is an enrichment of carriers at the interface.

Example: Suppose we are making a Ti contact to an unrealistically, lightly doped GaAs doped 10^{12} *n*-type. Draw the junction and sketch the charge distribution (GaAs ($n = 10^{12}/\text{cm}^3$, $X = 4.07$ eV) to Ti ($\Phi_m = 4.33$ eV).

Solution: Following the example of Sect. 6.7, the Fermi level is located 0.3 eV above the intrinsic Fermi level and 0.42 eV below the conduction band, as illustrated below.



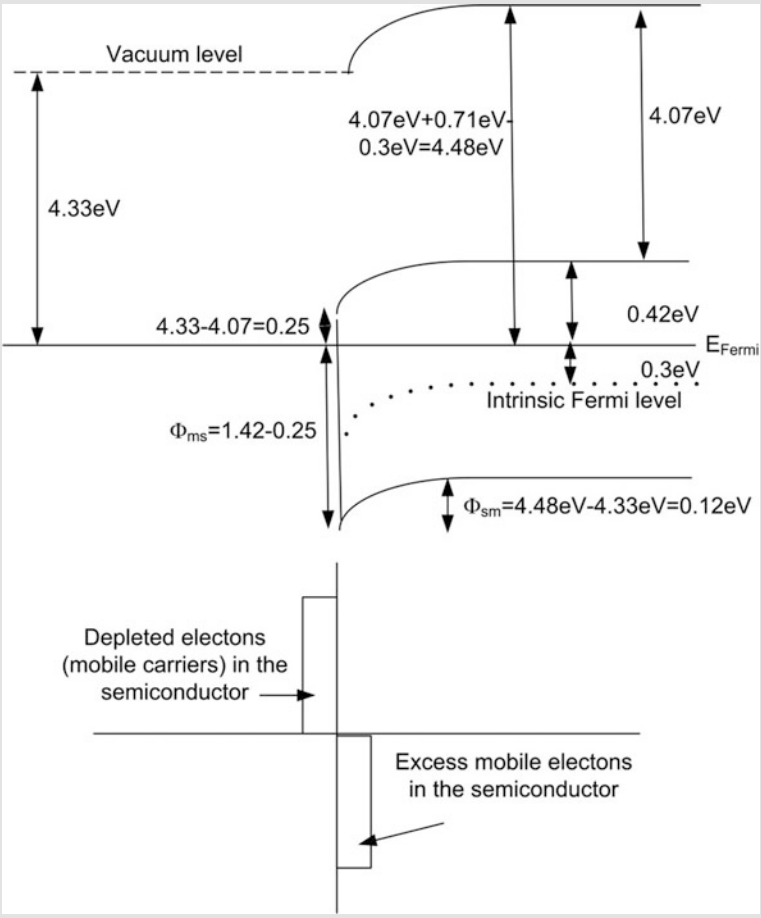
The junction is exactly the same as it was, except that in this case the majority carriers are electrons, and so the barrier to majority carries is 0.25 eV.

$$\Delta E_{n\,\text{metal}\to\text{semi}} = \Phi_{ms} = q\phi_m - qX = 4.33 - 4.07 = 0.25\,\text{eV}.$$

The work function for the semiconductor is 4.07 eV + 0.41 eV, or 4.48 eV. The degree of bending of the semiconductor bands is given by,

$$q\Phi_{sm} = q(\Phi_m - \Phi_s) = 4.33 - 4.48 = -0.15\,\text{eV}$$

The bands bend down 0.15 eV. However, if the majority carriers are electrons, the bands bending down (toward the Fermi level) actually mean an enrichment of carriers at the junction (more electrons than in the bulk). Hence, there is no barrier to electron flow from the semiconductor to the metal. This junction has no depletion layer; instead, it has excess mobile charge. Putting it together, the band structure and the charge density implied by it are given below.



This junction does not have an exponential *I–V* curve. Instead, it has an ohmic *I–V* curve. So what is wrong with this contact?

The first thing is that that level of semiconductor doping is not very conductive. In order to conduct carriers to the active region, the semiconductor should have relatively low resistance, hence, high doping.

It turns out that with most semiconductors and available metals, it is impossible to get a classic ohmic contact, for the following reason. Assume the semiconductor has to be heavily doped. In that case, the possible values of the work function are (roughly) either the electron affinity (for *n*-doped semiconductors) or the electron affinity plus the band gap for *p*-doped semiconductors.

Table 6.3 Some values of metal work functions and values of semiconductor work functions for *n*- and *p*-doped semiconductors. For a good *n*-ohmic contact, the work function of the metal should be less than that of the semiconductor; for a good *p*-ohmic contact, the metal work function should be greater

Metal (Φ_m)	Highly *n*-doped semiconductor work functions	Highly *p*-doped semiconductor work functions
	GaAs (4.07)	
Ti 4.33 eV		
	InP (4.35)	
Be 4.98 eV		
Au 5.1 eV		
Ni 5.15 eV		
		GaAs(5.49)
		InP (5.62)
Pt 5.65 eV		

For an *n*-doped semiconductor to bend down to form an ohmic contact, the work function of the semiconductor has to be greater than that of the metal. Most useful metals have work functions greater than 4.3 eV; typical semiconductors have electron affinities less than 4.3 eV. Table 6.3 illustrates this point by showing the work function of some metals and the potential work functions of doped GaAs and InP.

The key point of this table is that it is difficult to get good metal contacts to lasers. There are not many metals that have a work function that is less than the semiconductor electron affinity or greater than the electron affinity plus the band gap. In the next section, we will talk about how ohmic contacts can be realized.

6.7 Realization of Ohmic Contacts for Lasers

In reality, what is usually done for lasers is to use the best metals possible. The metal Pt is usually used for the *p*-metal contact, because of its high work function. Contact to the *n*-side is frequently made with NiGe alloys or Ti (both relatively low work function metals).

Schottky metal–semiconductor junction theory, as presented here, is partially an approximation. It is a guideline to conduction behavior across the junction, but not the whole story. Junction theory ignores the fact that the band structure at the surface of the semiconductor (where the metal is deposited) is different than in the bulk of the semiconductor. The surface has dangling bonds which tend to pin the Fermi level in the middle of the band gap.

To understand how we actually get good, low-resistance ohmic contacts, let us look at mechanism for current conduction through a metal–semiconductors junction.

6.7.1 Current Conduction Through a Metal–Semiconductor Junction: Thermionic Emission

Let us look first at the *I–V* equation for a Schottky junction and the methods for current conduction. In a Schottky junction, for current to get from the semiconductor to the metal side, it has to get over the potential energy barrier Φ_{sm} indicated. That barrier is a function of applied voltage. The figure shows that some carriers from the semiconductor manage to make it over the barrier onto the metal side, and at the same time, some carriers from the metal side manage to make it over the semiconductor side. In equilibrium, of course, these are equal, and there is no net charge flow.

Figure 6.10 (left) shows a Schottky junction in equilibrium, with the metal–semiconductor and semiconductor–metal contacts equal. The middle picture shows the junction with an applied forward bias. The barrier from semiconductor to metal side is lowered, and so the charge flow from semiconductor to metal side is increased.

The rightmost picture of Fig. 6.10 shows the junction with a reverse bias. In this case, the barrier on the semiconductor side is increased, and the charge flow from semiconductor to metal is decreased. (Apologies for confusing the reader: Schottky junctions are majority carrier conductors, and so charge transfer of electrons from the *n*-side to the metal corresponds to current flow in the opposite direction. We use 'charge flow' instead of current in this section to avoid this confusion.)

We note that regardless of bias, the charge flow from metal to semiconductor (limited by the barrier Φ_{ms}) stays about the same. This is analogous to the drift current flow in a *p–n* junction, which is also independent of applied bias.

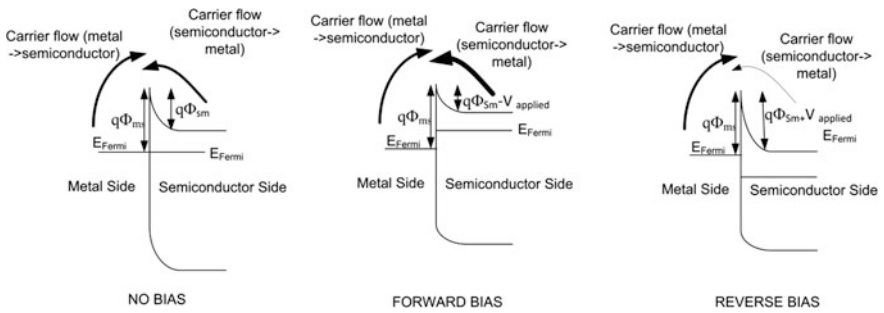


Fig. 6.10 Band structure of Schottky junction, under equilibrium, forward bias, and reverse bias

This method of current flowing through a Schottky junction is called thermionic emission. Even though there is a barrier for charge on the semiconductor to go over, because of the Fermi function and the nonzero temperatures, some carriers in the semiconductor will have an energy higher than that of the barrier, and it will be those that get conducted over the top.

Very qualitatively, the number of carriers at an energy sufficiently high to get over the barrier is exponentially dependent on the voltage. Therefore, roughly, the *I–V* curve of a Schottky junction looks like,

$$I = I_0(\exp(qV/kT) - 1) \tag{6.35}$$

In this book, we will not go any further into the saturation current I_0, but it depends on the details of the junction in ways similar to *p–n* junctions.

6.7.2 Current Conduction Through a Metal–Semiconductor Junction: Tunneling Current

There is another conduction mechanism that is possible for Schottky junctions that is not possible in *p–n* junctions. Examine the band diagram below. There are many states close to the carriers in the conduction band of the semiconductor on the metal side, separated only by the barrier. If the carriers can tunnel through the barrier, current can be conducted that way, as shown in Fig. 6.11.

This is the reason that the contact layers in semiconductors are very highly doped. The more highly doped, the thinner the depletion layer turns out to be. A thin depletion layer facilitates tunneling current. If the 'barrier' is thin enough, quantum mechanics allows current to go through it.

Another key to getting a good ohmic contact is annealing the contact after the metal is deposited. Typically, semiconductor wafers are heated to 400–450 °C after they are fabricated, for the purpose of encouraging some diffusion of the metal

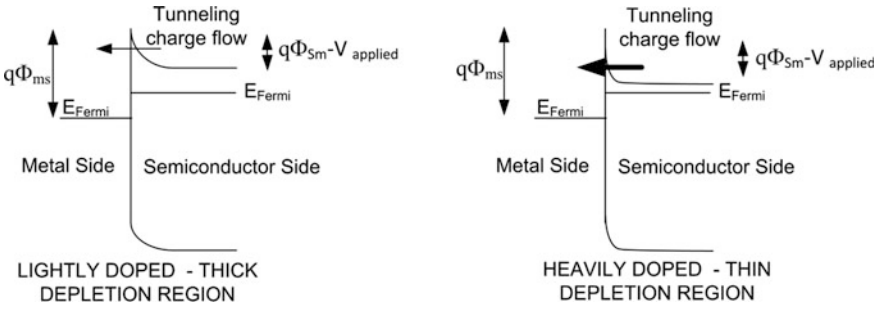


Fig. 6.11 Tunneling current through the depletion region of a Schottky barrier. Because the depletion region is thinner in a more highly doped semiconductor, having a highly doped semiconductor region facilitates tunneling current

atoms into the semiconductor. This junction is not the abrupt Schottky junction pictured, but it facilitates conduction and is quite important to device fabrication.

6.7.3 Diode Resistance and Measurement of Contact Resistance

Before we leave this metal–semiconductor junction topic, we should talk briefly about the resistances in a laser diode. Figure 6.12 has a schematic diagram of a ridge waveguide laser diode, showing the active region in the middle, the cladding on the *p*- and *n*-side, and the metal contact. Typical dimensions of a ridge waveguide laser are indicated. The resistance measured comes from both the contact resistance (associated with the metal–semiconductor junction) and a semiconductor conduction resistance, through the cladding regions.

The resistance, R_{semi}, of the semiconductor region, as a function of the geometry, is

$$R_{\text{semi}} = \frac{\rho l}{A}, \tag{6.36}$$

where A is the cross-sectional area through which the current flows, and l is the length of the region. The resistivity, ρ, depends on the doping and the material and is given by,

$$\rho = \frac{1}{q\mu_{n/p}N_{\text{D/A}}} \tag{6.37}$$

where N and μ are the appropriate doping density in the semiconductor and mobility, respectively.

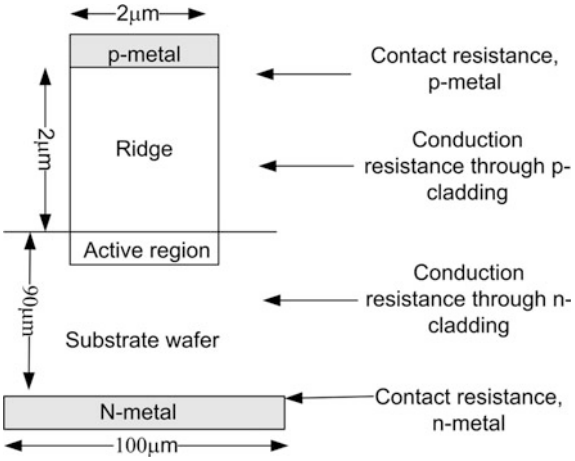


Fig. 6.12 A typical semiconductor ridge waveguide laser, showing the origins of the resistance terms including contact resistance between the semiconductor and metal, and the conduction resistance through the semiconductor

To give a sense of the relative importance of the various terms, look at the example below.

Example: In Fig. 6.12, the doping density of the ridge, and of the substrate, is 10^{17} cm^3, and it is 2 μm high, 2 μm long, and 300 μm long. The thickness of the wafer is 90 μm. Find the resistance due to the top and bottom cladding regions (μ_n is 4000 cm^2/V s, and μ_p is 200 cm^2/V s).

Solution: Because the bottom region is very large, the cross-sectional area is quite large. Typically, the bottom *n*-metal can be 100 μm wide or larger. Taking the average of 100 μm- and the 2 μm-wide active regions gives a 50 μm-wide bottom region.

The top region is much more constrained and is only 2 μm wide.

The resistivity associated with the *n*-region is therefore $1/(1.6\times10^{-19})(4000)10^{17} = 0.016\,\Omega\,\text{cm}$, and the resistivity associated with the *p*-region is 20 times greater (0.31 Ω cm) due to the 20× lower mobility.

The resistance of the *n*-contact region is about $\frac{0.016(90\times10^{-4})}{50\times10^{-4}(300\times10^{-4})} = 1\ \Omega$. The resistance of the *p*-contact region is much higher $\frac{0.31(2\times10^{-4})}{2\times10^{-4}(300\times10^{-4})}$, or about 10 Ω.

This is typical of lasers, where much of the resistance is in the *p*-cladding. Typically, the undoped regions near the active region are insignificant, because they are so thin; the highly doped contact layers are also insignificant, because they are highly doped. It is the moderately doped cladding which adds most of the resistance.

Typical specified values for laser resistances are less than 8 Ω for directly modulated devices.

The contact resistance associated with the metal–semiconductor junction can be experimentally measured with a lithographic pattern, as shown below. The measurement of each pair of pads includes two contact resistances plus the semiconductor resistance. Measurements of a few resistances versus length will extrapolate to twice the contact resistance as shown in Fig. 6.13.

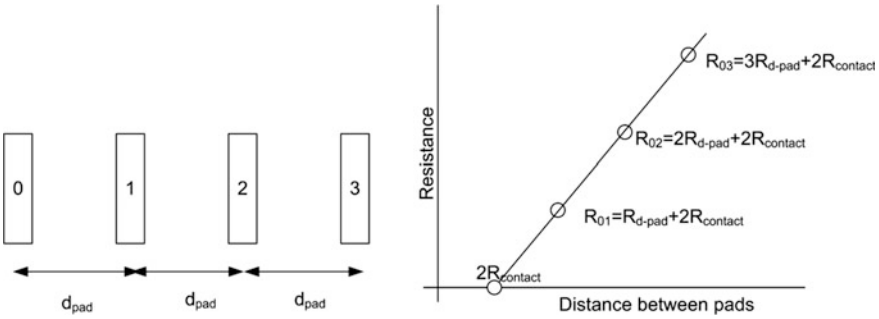


Fig. 6.13 Left, metal pads on a semiconductor with fixed spacing; right, measurement of resistance between pairs of pads. Extrapolated to zero length, it gives twice the contact resistance

6.8 Summary and Learning Points

In this chapter, the details involved with injecting current into the active region are described, including the similarities and differences between laser diodes and standard diodes, and the details of making good metal contact to semiconductors.

A. The electrical characteristics of semiconductor lasers are also important to their operation. Low-resistance contacts lead to lower ohmic heating.
B. Semiconductor lasers are fundamentally *p*–*n* junctions.
C. The *p*–*n* junctions form a depletion region, where the mobile electrons and holes recombine and leave behind immobile depletion charge.
D. The depletion charge gives rise to an electric field and a built-in voltage between one side and the other side of the junction.
E. On each side of the depletion region is what is called the quasi-neutral region, where the net charge is zero.
F. The boundaries between the depletion region and the quasi-neutral region are assumed to be abrupt.
G. The electric field across the depletion region gives rise to a drift current, going from the *n*-side to the *p*-side; in addition, there is a diffusion current, going from the *p*-side to the *n*-side. These currents are balanced in equilibrium.
H. Applied forward bias reduces the built-in voltage. The magnitude of the drift current remains approximately the same, but the magnitude of the diffusion current increases exponentially.
I. Assuming an abrupt junction and a Fermi level split across the junction, the number of excess carriers injected into each side of the quasi-neutral region depends exponentially on voltage.
J. These excess carriers recombine as they diffuse into the quasi-neutral region.
K. From this diffusion/recombination process, the diode *I*–*V* curve showing in Fig. 6.8 can be derived.

L. Lasers differ significantly from *p–n* junctions.
M. Lasers have significant recombination current, and so the diode ideality factor is typically closer to two than one.
N. Above threshold, the quasi-Fermi level in lasers is clamped. Hence, the excess carriers do not increase the carrier density in the quasi-neutral region but instead increase the number of photons out.
O. This gives rise to a constant differential resistance above threshold; the exponential *I–V* curve is no longer followed.
P. The general problem of making metal contacts to semiconductors is described by Schottky theory.
Q. Assuming the band structure of the semiconductor is the same at the surface as in the bulk, the band diagram can be drawn by drawing a constant Fermi level and a continuous vacuum level. This gives rise to band banding in the semiconductor.
R. This band bending represents the depletion region (if the band bends away from the Fermi level) or carrier enhancement (if the band bends toward the Fermi level)
S. The balancing charges accumulate on the metal side.
T. An applied bias reduces the barrier on the semiconductor side, since the barrier on the metal side is fixed by the material constants.
U. To obtain an ohmic contact, the work function has to be less than the electron affinity (for *n*-doped semiconductors) or greater than the electron affinity plus the band gap (for *p*-doped semiconductors).
V. Practically speaking, the work functions of most metals do not satisfy condition (B); therefore, usually, the contact to a semiconductor is not a perfect ohmic contact.
W. It works as an ohmic contact because (a) the band structure at the surface is usually different than in the bulk, (b) the surface is heavily doped to make the depletion layers thinner, and (c) the contact is annealed, to blur the junction further.
X. The annealing is very important to semiconductor laser operation.
Y. Typically, semiconductor resistances derive from conduction resistance through the *p*-cladding and metal semiconductor contact. They are usually specified to be 8 Ω or less.

6.9 Questions

Q6.1. If the current conduction across the depletion region is drift and diffusion, and near the junction in the quasi-neutral region is diffusion only, how does current get from the contacts to the junction?
Q6.2. Would you expect there to be a generation, or a depletion term, in general in the semiconductor depletion region?

Q6.3. Annealing usually improves the semiconductor–metal interface, lowering the resistance and making it more ohmic. Can you think of some potential problems with over-annealing?
Q6.4. Why is Eq. 6.15 true?

6.10 Problems

P6.1. An InP semiconductor is *p*-doped to $10^{18}/cm^3$. Find the Fermi level and the concentration of holes and electrons in the semiconductor.
P6.2. The sample in P6.1 is illuminated with light, such that 10^{19} electron–hole pairs are created per second per cm^3. The lifetime of each electron or hole is 1nS.

(a) Is the semiconductor in equilibrium?
(b) What is the steady state value of excess electrons and holes in the semiconductor (this is equal to the generation rate multiplied by the lifetime).
(c) What is the quasi-Fermi level of electrons, and holes, now in the semiconductor?
(d) Compare the location of the Fermi level in P6.1 with the location of the quasi-Fermi levels calculated here. Between the holes and the electrons, which shifted more and why?

P6.3. A semiconductor GaAs *p–n* junction has the following specifications:

p-side	I	*n*-side
$N_A = 5x10^{17}/cm^3$		$N_D = 10^{17}/cm^3$
$\tau_n = 5$ μs		$\tau_p = 10$ μs
$\mu_p = 400$ cm^2/V-s		$\mu_p = 350$ cm^2/V-s
$\mu_n = 8000$ cm^2/V-s		$\mu_n = 7500$ cm^2/V-s

(a) Sketch the band structure and calculate V_{bi}.
(b) Calculate the depletion layer width.
(c) Calculate the peak electric field in the depletion region.
(d) Calculate the forward current under 0.4 V applied bias in A/cm^2.
(e) Why is the mobility of holes and electrons slightly less on the *p*-side?
(f) Assume the *p–n* junction above is actually a laser, which has an additional undoped region 3000 Å wide between the *p*- and the *n*-region. Roughly, estimate the peak electric field in the *i* region.

P6.4. A sample of GaAs is linearly doped with N_D going from 10^{14} to $10^{17}/cm^3$ over 1 mm.

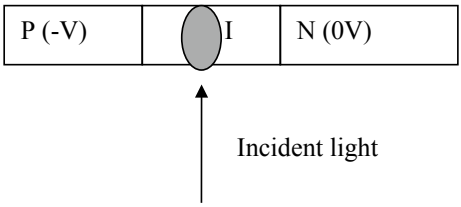


Fig. P6.14 A *p–i–n* diode with a small pulse of incident light that creates excess holes and electrons

(a) Sketch the band diagram of the sample, indicating the conduction band, the valence band, the Fermi level, and the intrinsic Fermi level.
(b) Indicate the kind and direction of the charge flow in the sample.
(c) Indicate the kind, and direction, of currents in the sample.
(d) Is there any fixed charge in this sample, and if so, where is it?

P6.5. A reverse-biased *p–i–n* GaAs-based photodetector has a light shined momentarily on it in the center of the i-region, creating a small region with excess holes and electrons (equivalent to moderately doped levels, 10^{16}/cm^3). The *p*- and *n*-regions are fairly heavily doped (10^{18}/cm^3) (Fig. P6.14).

(a) Ignoring the excess holes and electrons created by the absorption of light, sketch the depleted regions of the semiconductor and indicate the direction of the electric field.
(b) Sketch the band diagram of the device *clearly labeling* the electron and hole quasi-Fermi levels and the applied voltage *V*. Include the effect of the excess optically created holes and electrons.
(c) Indicate the direction in which the excess holes and electrons created by the light pulse will travel.
(d) Assume now that the diode is moderately *forward* biased, and a brief pulse of light is again shone in the center of the *i* region.
(e) Sketch the band diagram of the device, indicating electron and hole quasi-Fermi levels and the applied voltage *V*. Indicate again the direction the excess holes and electrons will travel.
(f) Assume the light is misaligned and now shines in the middle of the *p*-region. Sketch the band diagram of the device indicating the electron and hole quasi-Fermi levels. Again, do not neglect the effect of the optically created holes and electrons.

P6.6. A Schottky barrier is formed between a metal having a work function of 4.3 eV and Si (Si has an electron affinity of 4.05 eV) that is acceptor doped to 10^{17}/cm^3.

(a) Draw the equilibrium band diagram, showing V_0 and ϕ_m.
(b) Draw the band diagram under (a) 0.5 V forward bias, (b) 2 V reverse bias.

P6.7. For the system used in Problem P6.6, what range of Si doping levels and types will give rise to an ohmic contact in Si?

P6.8. Derive an equation for the work function of a *p*-doped semiconductor in terms of doping and its material parameters.

P6.9. Draw the band diagram of an n–n^+ semiconductor junction in equilibrium. Label the electric field (if there is one), the drift current (if there is drift current) and the diffusion current (if there is diffusion current).

P6.10. In Fig. 6.12 and the associated example, find the doping necessary to reduce the top cladding resistance to 5 Ω.

The Optical Cavity

7

Macavity, Macavity, there's no one like Macavity,
There never was a Cat of such deceitfulness and suavity.
—T.S. Eliot, Old Possums Book of Practical Cats

Abstract

In this chapter, the design and characteristics of a typical semiconductor laser optical cavity are examined. The concept of free spectral range and single longitudinal and spatial modes are defined, and procedures for designing single-mode optical cavities are discussed.

7.1 Introduction

In this book, we began by talking about the general properties of lasers and determined that the requirements for a laser were a non-equilibrium system with high optical gain and a high photon density. In subsequent chapters, we focused on the first requirement for a high optical gain, and the various constraints, limits, and considerations in getting the necessary high gain at the correct wavelength from a semiconductor active region.

Now, we would like to turn our attention to the second requirement of a high photon density. This high photon density is achieved by putting the gain region into a cavity which holds most of the photons inside. For the HeNe gas laser discussed in Chap. 2, the cavity is simply a pair of mirrors at each end of a laser tube. For the semiconductor lasers we discuss now, this optical cavity is a dielectric waveguide formed by the geometry of the laser and the index contrast between the layers within the laser. A good laser is a good waveguide. This laser property is so important that this entire chapter is devoted to waveguides, in general, with special attention paid to common laser waveguide types.

© Springer Nature Switzerland AG 2020

D. J. Klotzkin, *Introduction to Semiconductor Lasers for Optical Communications*,
https://doi.org/10.1007/978-3-030-24501-6_7

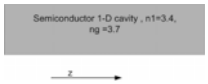
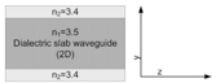
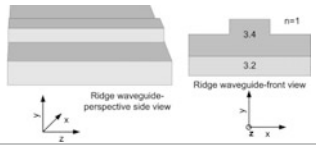
The simplest semiconductor laser cavity is a cleaved piece of semiconductor (typically a few hundred microns long). This cavity type defines a Fabry-Perot laser: the cleaves, which are close to atomically smooth, act as excellent dielectric mirrors and can keep the photon density within the cavity high. Even, this very simple cavity profoundly affects the light generated in the cavity.

In practice, there are many other cavities which are used, including vertical Bragg reflectors, integrated distributed feedback lasers, and even devices based on total internal reflection. In this chapter, we are going to focus on the effect of the cavity on the light, and particularly the design of the optical cavity to realize the desired single-mode characteristics.

7.2 Chapter Outline

We are going to navigate systematically from a one-dimensional picture, in which we consider only the direction of propagation of the light, to a two- and three-dimensional picture, in which we consider the direction of propagation of light, and the one and two dimensions transverse to it in order to get a full picture of the influence of the optical cavity on the emitted light. Table 7.1 is intended to aid the reader in navigation. It outlines completely the kind of optical cavity that we are looking at and the learning point we are trying to illustrate for the reader.

Table 7.1 Types of optical structures considered, their appropriate section, and the learning point intended from each

Type of structure	Picture with coordinate system	Learning point	Section (s)
Pair of reflecting mirrors (etalon) in air		Effects of cavity length on longitudinal mode (wavelength) spacing and supported wavelengths	7.4.1
Dielectric sandwiched by air		Effect of cavity group index on longitudinal mode (wavelength) spacing	7.4.3
Two-dimensional slab waveguide		Influence of dielectric thickness and index on spatial mode properties in 2D	7.5
Three-dimensional ridge waveguide		Influence of dielectric thickness and index on spatial mode properties in 3D	7.6

Let us make one important distinction here, and we will return to it the appropriate sections. The word 'mode' in a laser context has several meanings. In Sect. 7.4, laser longitudinal mode means the allowed wavelengths in the cavity. A gain region emitting around 1300 nm placed in the optical cavity of a laser will emit specific wavelengths associated with specific longitudinal modes (for example, 1301.2, 1301.8 nm, and more).

Section 7.5 focuses on the transverse distribution of the light of a particular wavelength within a cavity. For example, if light of a specific wavelength is traveling in the *z*-direction, the optical field distribution in the *y*-direction could have one spatial mode showing a single optical field peak in the center of the waveguide, and a second one with two peaks (for a multimode waveguide).

Mode can also refer to the polarization state (as in 'transverse electric' or 'transverse magnetic' mode.). The meaning is usually clear from the context. Each of these types of modes will be revisited in their associated sections.

7.3 Overview of a Fabry-Perot Optical Cavity

Figure 7.1 shows a picture of the laser emphasizing its optical cavity and waveguide qualities. This common laser cavity is called a ridge waveguide Fabry-Perot. The cavity is formed by a laser bar cleaved from a wafer forming two cleaved

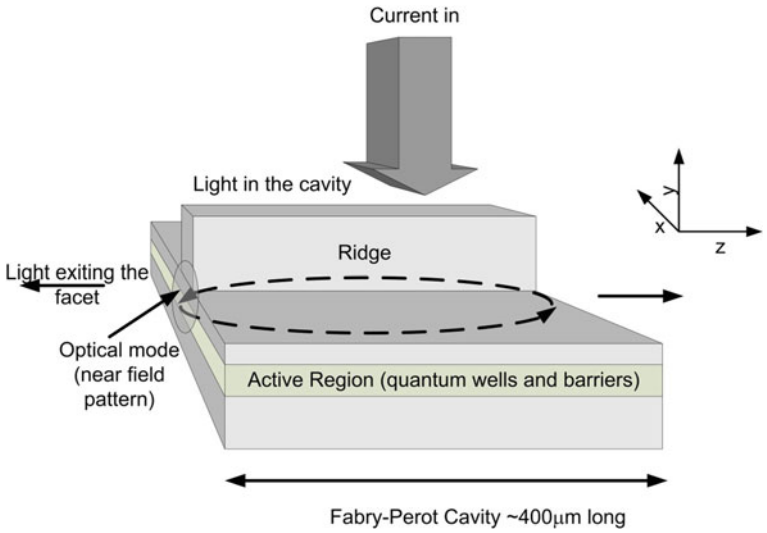


Fig. 7.1 A picture of a Fabry-Perot cavity (ridge waveguide) structure, showing light bouncing back and forth between the two facets with light exiting the facets at each end. Qualitatively, the presence of the ridge confines the ridge in the *x*-direction, the index contrast in the active region confines the light in the *y*-direction, and the optical mode bounces back and forth between the facets in the *z*-direction

semiconductor facets, with current injected through the top and bottom, and light emitted from the front and back. This edge-emitting device is the simplest optical cavity to realize; this structure is used commercially, usually with the cleaved facets coated to enhance or reduce reflectivity.

The light in the laser cavity bounces back and forth between the two facets in the z-direction while it is confined in the waveguide formed by the laser. Qualitatively, the higher index of the quantum wells (compared to the surrounded layers) confines the light in the y-direction, and the presence of the ridge above the quantum wells confines the light in the x-direction. The reflection back and forth in the z-direction results in only certain, regularly spaced wavelengths in the cavity (called *free spectral range*), and the confinement in x-y affects the intensity pattern (the *lateral or spatial mode shape*) of the light in the laser. This overview is intended to put that discussion of free spectral range and optical modes to follow into the proper laser context.

Figure 7.1 shows a combined view of the semiconductor active region serving as the optical cavity. A view of the device solely as an optical cavity is shown in Fig. 5.1.

7.4 Longitudinal Optical Modes Supported by a Laser Cavity

7.4.1 Optical Modes Supported by an Etalon: The Laser Cavity in 1D

First, let us look at the cavity in strictly one-dimensional view as light between a pair of mirrors. Optical plane waves emanate from it originating from the recombination (stimulated or spontaneous) of carriers within the cavity. Let us consider the optical wavelengths supported by the cavity in Fig. 7.1 and think of the light as strictly a wave phenomenon.

Imagine spontaneous emission light of a range of wavelengths being created within the cavity and then bouncing back and forth between the mirrors. In order for any given wavelength to be allowed in the cavity, the round trip light has to undergo constructive interference. Mathematically, a round trip for any given wavelength has to be an integral number of wavelengths. Equation 7.1 states this succinctly.

$$m = \frac{2L}{(\lambda/n)} = \frac{2Ln}{\lambda} \tag{7.1}$$

This idea is illustrated in Fig. 7.2.

Figure 7.2 shows a set of cavities sandwiched by two reflective mirrors. Because of the coherent nature of light, only certain wavelengths are supported in any cavity, depending on the length of the cavity and the wavelength. In this set of figures (a–f),

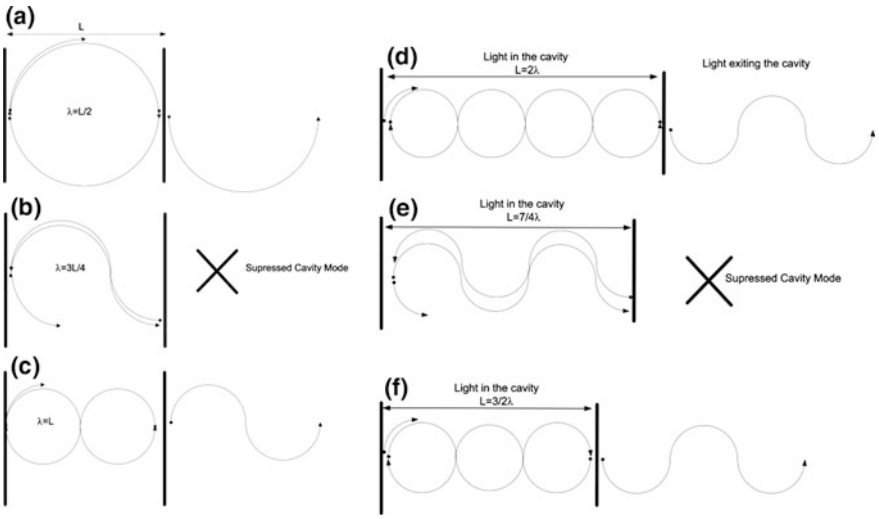


Fig. 7.2 (**a**–**c**) show several optical wavelengths in the same length of cavity (right) and the same optical wavelength in three different cavity lengths (left), illustrating how the interaction of the cavity and the wavelength create supported and suppressed cavity modes

the actual peaks and valleys of the optical wave represent the phase of the light; the peaks and valleys represent the change in phase as it propagates, and so the distance between two peaks (or valleys) is the wavelength.

Figures 7.2a–c show three different wavelengths in one optical cavity. In Fig. 7.2a, the optical cavity is exactly half a wavelength, so the round trip (of one wavelength) supports constructive interference. Figure 7.2b shows a cavity that is three-fourths of a wavelength long, so the round trip is one-and-a-half cavity lengths long. After one round trip, the original light is out of phase by 180°, and so this cavity cannot support this wavelength. Figure 7.2c shows wavelength equal to the cavity length.

Figures 7.2d–f illustrate the same idea, with the same wavelength shown in three different size cavities. The first cavity (Fig. 7.2d) is exactly 2λ of the light long. As the light travels one round trip, it comes back to the mirror and is reflected again, exactly in phase with where it started. Since this particular wavelength is constructively interfered with in the cavity, this wavelength is supported in this cavity.

The cavity shown in Fig. 7.2e is 7/4λ of a wavelength long. The round drive is three and a half wavelengths which results in this wavelength being 180° out of phase with itself and not being supported. The cavity of Fig. 7.2f is 3/2λ and supports that wavelength.

Just as the net gain has to be 1 in order for the laser to be in steady state, the net phase, for a round trip, has to be a multiple of 2π. For a laser above threshold, Eqs. 7.1 and 5.3 can be combined into a single equation, as

$$R_1 R_2 e^{(g+jk)2L} = R_1 R_2 e^{\left(g+j\frac{2\pi}{\lambda/n}\right)2L} = 1 \tag{7.2}$$

where g is the gain, k is the propagation constant $2\pi/\lambda$ in the cavity, n is the cavity index, L is the cavity length, and R_1 and R_2 are the facet reflectivities.

7.4.2 Free Spectral Range in a Long Etalon

Qualitatively, the idea of interference of coherent light leads to a set of 'allowed' optical wavelengths supported by the cavity, and 'forbidden' optical wavelengths that the cavity does not support. In this section, let us define the standard optical terminology that is used to specify etalons, and then in the next section discuss what this means for the spectrum of Fabry-Perot lasers.

A very simple cavity is composed of two mirrors spaced a distance L apart and is illustrated in Fig. 7.3. The index of this pedagogical cavity is assumed to be wavelength-independent and equal to 1. Let us consider the optical wavelengths, and the wavelength spacing allowed by the cavity. In this example, the cavity length is 1 mm, much longer than the optical wavelength.

The modes supported by such a cavity are qualitatively shown in Fig. 7.3, with the spacing between them defined as the *free spectral range* (FSR). With a long cavity, the modes will be closely spaced, as described in Eq. 7.1 and in a free spectral range equation to be derived below.

A good qualitative way to understand Eq. 7.1 is that in a cavity with reflection from the facets, the round trip path length *2L* has to be an integral number of wavelengths in the cavity. In the cavity shown (1 mm long), a wavelength of 1600 nm will have an integral number of 1250 wavelengths in a round trip between the mirrors.

A slightly shorter wavelength with 1251 wavelengths of light in a round trip is also supported by this cavity. That wavelength is 2 mm/1251, or 1598.7 nm. For each integral number that the number of wavelengths in the cavity is incremented,

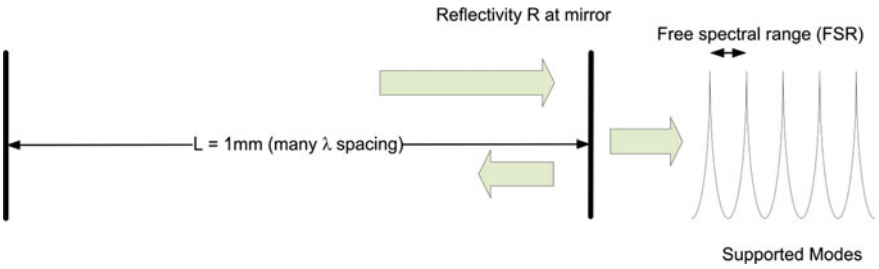


Fig. 7.3 An optical cavity composed of air sandwiched by two reflective mirrors which supports a number of optical modes separated by the free spectral range (FSR). In this picture, the optical cavity is presumed to be many wavelengths long, and in air, with an index of $n = 1$

there will be another allowed wavelength. In this example, the spacing between them, or free spectral range, is 1.3 nm.

Example: Calculate the next higher wavelength supported by the cavity shown in Fig. 7.3 with a length of 1 mm.

Solution: The next higher wavelength will have one fewer full wavelength in a round trip through the cavity, or 1249. Two mm/1249 is 1601.3 nm.

Example: Calculate the free spectral range of this cavity.

Solution: From simply examining the space between peaks, the free spectral range is about 1.3 nm. We will derive an expression for it below.

Let us develop an expression for the free spectral range which measures the spacing between the peaks. We will start by labeling λ_m the wavelength associated with m round trips through the cavity, and λ_{m+1} the slightly shorter wavelength associated with $m + 1$ round trips through the cavity. The requirement for an integral number of wavelengths in a round trip is

$$2L = \frac{m\lambda_m}{n} = \frac{(m+1)\lambda_{m+1}}{n} \tag{7.3}$$

from which we can write

$$\frac{m\lambda_m - (m+1)\lambda_{m+1}}{2Ln} = 0 \tag{7.4}$$

or

$$m\Delta\lambda = \lambda_{m+1} \tag{7.5}$$

This expression, while correct, is not satisfying since it requires a calculation for m (the number of round trips). It can be shown (see Problem P7.1) by substituting for m that the free spectral range is

$$\Delta\lambda \approx \frac{\lambda_{m+1}^2}{2Ln} \tag{7.6}$$

Equation 7.6 gives the spacing of the modes, $\Delta\lambda$, as a function of the index and the cavity length. The important point is that mode spacing depends inversely on the length of the cavity, and the cavity index, and directly on the central wavelength squared.

7.4.3 Free Spectral Range in a Fabry-Perot Laser Cavity

A Fabry-Perot laser cavity has some important differences from the mirrored etalon described above. In its simplest model, shown in Fig. 7.4, below, it is a smooth piece of dielectric material with facet reflectivity due to the index contrast between the material and surrounding air. Unlike the sandwiching mirrors pictured in Figs. 7.2 and 7.3, the mirrors of this cavity are due to the index difference between the ambient atmosphere and the semiconductor, with the reflectivity given by Eq. 5.2.

More importantly, the wavelengths-of-interest of a laser active region are right around the band gap of the semiconductor. As shown in Fig. 7.5, around the band gap, the refractive index and gain are very dependent on wavelength. Because of this strong dependence of refractive index, the equations for free spectral range will turn out to be slightly modified in a semiconductor laser.

If the index for two wavelengths λ_m and λ_{m+1} is slightly different, like Fig. 7.5 says, we can rewrite Eq. 7.3 as

$$2L = \frac{m\lambda_m}{n_m} = \frac{(m+1)\lambda_{m+1}}{n_{m+1}} \tag{7.7}$$

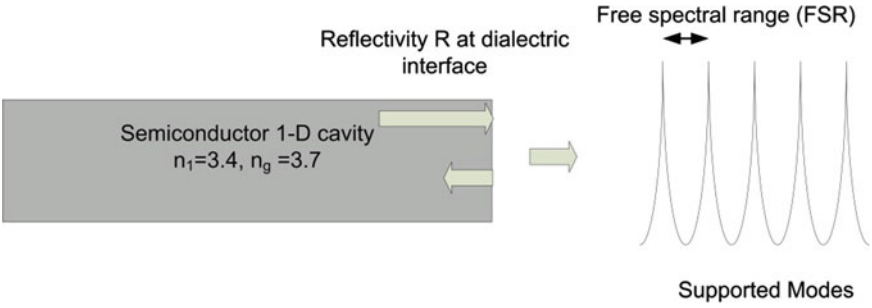


Fig. 7.4 A one-dimensional model of a dielectric cavity. The difference in index between the cavity and air provides the mirror, and the group index sets the spacing of the modes

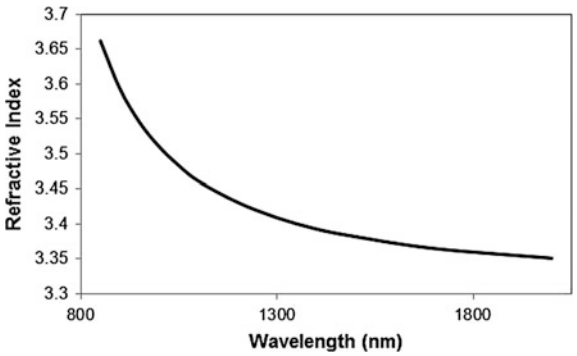


Fig. 7.5 Refractive index of GaAs at room temperature around its bandgap of ~870 nm at 300 K Adapted from http://www.batop.com/information/n_GaAs.html and data *in Journal of Applied Physics, D*. Marple, V. 35, pp. 1241

It can be shown (see Problem P7.1) that this expression leads to the following expression for free spectral range,

$$\Delta\lambda = \frac{\lambda_{m+1}^2}{2Ln_g} \tag{7.8}$$

where n_g is the group index, defined by

$$n_g = n - \lambda\frac{\Delta n}{\Delta\lambda} = n - \lambda\frac{\mathrm{d}n}{\mathrm{d}\lambda} \tag{7.9}$$

The group index captures both the index, and the change in index versus wavelength. Since the calculation of the mode spacing is based on a net 2π phase difference between two wavelengths covering the same length, this is the appropriate index to use.

However, the actual number of whole wavelengths in the cavity is given by the mode index, n. This subtle difference is illustrated in the example below.

Example: A 300-μm-long laser cavity has a mode index of 3.4191, a group index of 3.6432, and a lasing wavelength of 1399.359 nm (the need for such precise numbers will become clear throughout the problem)
Find the spacing of the cavity modes and the integral number of wavelengths in a round trip in the cavity. Find the next longer wavelength and estimate its mode index and the number of round trips in the cavity associated with that wavelength.

Solution:
From above, we can write

$$\Delta\lambda = \frac{\lambda^2}{2Ln_g} = \frac{1.399359^2}{2(300)3.6432} = 0.895834 \times 10^{-3}\ \mu\text{m}.$$

The spacing between peaks (or free spectral range) is about 0.9 nm. On the other hand, the integral number of wavelengths in the cavity is $2L/(\lambda/n)$, or 600 μm/(1.399359/3.4191) = 1466 wavelengths exactly.
The next longer wavelength is $1.399359 + 0.895834 \times 10^{-3}$ μm, or 1.400255 μm. The mode index of the next longer wavelength ($m = 1465$) is estimated as follows.

$$n_g = n - \lambda\frac{\Delta n}{\Delta\lambda} = 3.6432 = 3.4191 - 1.399353\frac{\Delta n}{\Delta\lambda} \text{ gives } \frac{\Delta n}{\Delta\lambda} = -0.16/\mu\text{m}.$$

Then, the mode index at 1.400255 (the next longer wavelength) is $3.6432 - 1.400255(0.16) = 3.418957$, and the number of round trips is, $600/(1.40025/3.418957) = 1465$, exactly. Notice that if we had used the same index for 1.399359 as for 1.400255, the calculated number of modes would have been $600/(1.40025/3.4191) = 1465.06$, a non-integral number. It is the slight shift in index between adjacent wavelengths that makes the condition of Eq. 7.1 work out exactly for each of the cavity wavelengths.

7.4.4 Optical Output of a Fabry-Perot Laser

With the idea that a Fabry-Perot optical cavity is an etalon, supporting a discrete set of wavelengths, let us take a look at the output of a Fabry-Perot laser. The important characteristic of a Fabry-Perot laser is that the reflectance does not depend on wavelength. All the wavelengths are reflected approximately equally.

This gives rise to the expected output spectra (graph of power vs. wavelength) of a Fabry-Perot cavity. The wavelengths are spaced approximately evenly according to Eq. 7.8. The predicted peaks are seen in the region over which the semiconductor has net gain and emits photons (called the *gain bandwidth* region).

A typical output spectra from a Fabry-Perot laser emitting when biased above threshold is shown in Fig. 7.6. There are a few prominent modes in a range from 1290 to 1305 nm. Looked at on a logarithmic scale, emission could probably be seen over a range of 40 nm, but 100 times lower in power than the peaks that are shown.

This figure is surprising if you think about it. According to the rate equation model, the carrier density and optical gain are clamped above threshold, and after that, injected current leads to increased optical output. Since the gain reaches the threshold gain at one particular wavelength first, it would be reasonable to think that the light at the single wavelength which is lasing at threshold increases, and the light at the other modes (which are driven by spontaneous emission) should remain the same since the carrier population is clamped. Hence, we would expect one dominant wavelength out.

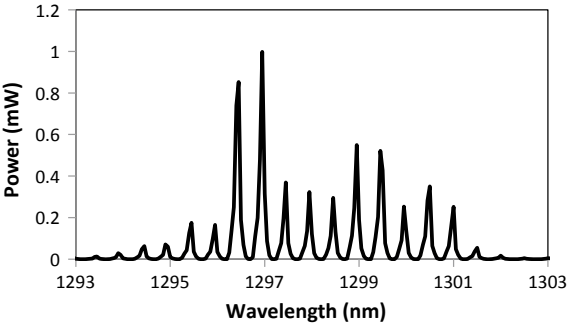


Fig. 7.6 Output spectrum of a Fabry-Perot laser

However, there are some non-ideal effects which make this simple model incorrect. In particular, there is a phenomenon called spectral hole burning. When a lot of light is produced at a specific wavelength, it reduces the gain at that wavelength and facilitates the production of light at other wavelengths. At high optical power levels, the carrier distribution is no longer accurately described by a Fermi distribution, which leads to lasing at more than one wavelength.

A phenomenological way to describe this is with the *gain bandwidth*, as a material property. The range of wavelengths over which lasing is supported is called the gain bandwidth (typically of the order of 10 nm or so) and the spacing of the modes in this gain bandwidth (determined by the cavity length) determines the number of lasing modes. The example below illustrates this idea.

Example: A particular material has a gain bandwidth of 15 nm at a lasing wavelength of 1.3 μm, a group index of 3.6, and an index of 3.4. In a cavity 250 μm long, about how many modes are lasing?

Solution: This is fairly straightforward. The spacing between cavity modes is

$$\Delta\lambda = \frac{\lambda^2}{2Ln_g} = \frac{1.3^2}{2(250)3.6} = 0.94\,\text{nm}.$$

The number of modes is about the gain bandwidth/mode spacing, or 16 modes. Note that as the cavity length increases, the mode spacing decreases and the number of distinct lines seen will increase as well.

7.4.5 Longitudinal Modes

Each of these lasing wavelengths which are within the gain bandwidth of the material is identified as the *longitudinal modes* of the devices. Each of these

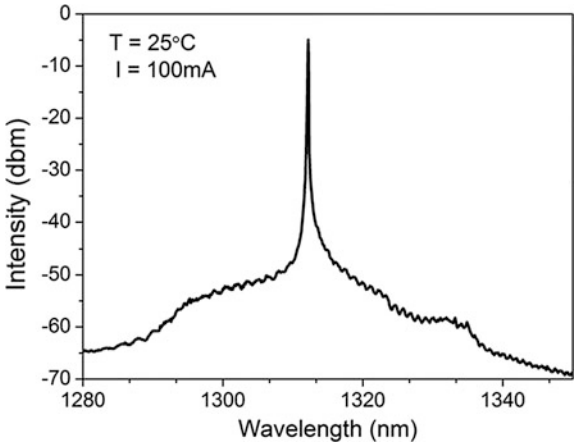


Fig. 7.7 Typical output spectrum of a distributed feedback (DFB) single longitudinal mode laser

wavelengths is associated with a different standing wave pattern in the cavity. For long-distance transmission, of course, a single wavelength with a single effective propagation velocity is required. For wavelength ranges that are not subject to dispersion (around 1300 nm) or low-cost solutions, Fabry-Perot lasers are sometimes commercially used, but in general, high-performance devices need to have only one wavelength.

These devices are almost universally distributed feedback lasers (DFBs) which will be discussed in-depth in a subsequent section. These DFBs have inherently low dispersion because they are single wavelength, and also have output wavelengths which are inherently less temperature-sensitive than Fabry-Perot. For multichannel wavelength division multiplexed (WDM) system, often single wavelength DFBs are required, not for dispersion but for wavelength stability over a specific temperature range.

While we are not yet going to explore the detailed fabrication and properties of DFB devices, for context and comparison, Fig. 7.7 shows a typical spectrum of such a device. Unlike the Fabry-Perot device in Fig. 7.5, it has only a single wavelength.

7.5 Calculation of Gain from Optical Spectrum

Now is an appropriate place to describe an experimental technique to measure the gain spectrum of a semiconductor laser. In Chap. 4, we discussed optical gain in terms of the density-of-state and injection level, and in Chap. 5, we showed that above threshold, the gain point of the active region cavity is actually set by the loss point of the cavity, which includes the absorption loss and the mirror loss.

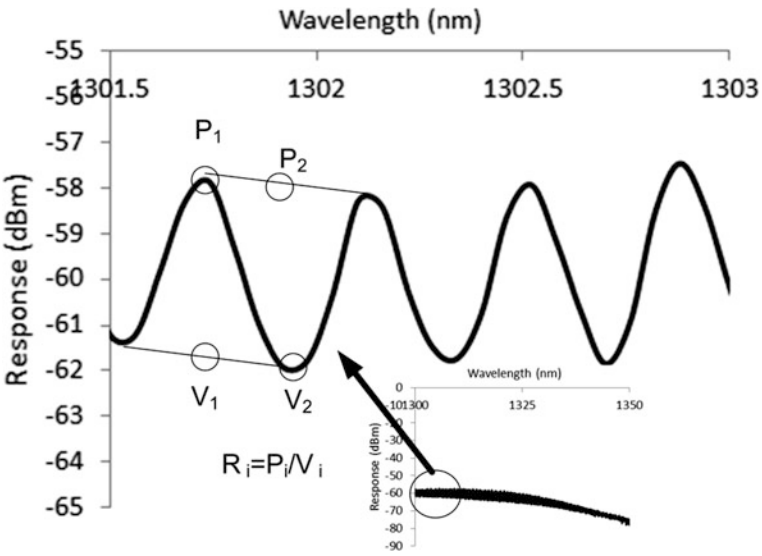


Fig. 7.8 A sub-threshold spectra, shown from 1300 to 1350 nm in the inset with a close-up view of the peaks and valleys from 1301.5 to 1303 nm in the main diagram

However, the below-threshold spectrum of the laser itself can tell you the net gain of the cavity, in the following way. As shown in Fig. 7.8, below threshold the light experiences gain as it travels within the cavity, but the gain is not quite enough to overcome the cavity loss. However, at some wavelengths, the light experiences constructive interference as it goes through the cavity (the peaks in the Fabry-Perot etalon spectrum) and at other wavelengths (the troughs at the Fabry-Perot spectrum), the light experiences destructive interference as it goes through the cavity. Hakki and Paoli[1] realized that actual gain spectra of the laser could be derived by looking at the ratio of the amplitude of the constructively interfered light to the destructively interfered light.

The process will be best illustrated by example. On the figure, we define a modulation index r_i as the ratio of the peak power to the valley power. Since the peaks and valleys do occur at different wavelengths, typically the 'peak' associated with a given valley is the average of the adjacent peaks, and the valley associated with a given peak the average of the adjacent peaks.

The net gain (or modal gain g_{modal}) is given by

$$g_{\text{net}} = g_{\text{modal}} + \alpha = \frac{1}{L}\ln\left(\frac{r_i^{1/2}+1}{r_i^{1/2}-1}\right) + \frac{1}{2L}\ln(R_1R_2) \tag{7.10}$$

[1]B. W. Hakki, T. L. Paoli, Journal of Applied Physics, vol. 46, pp. 1299, 1975.

where r_i is the ratio of peaks and valleys, as defined in the figure; L is the cavity length, R_1 and R_2 are the facet reflectivities of both facets, and α is the absorption loss in the cavity. (We note the form above is slightly different than the original Hakki-Paoli formulation, which omitted α and interpreted modal gain as optical gain plus absorption loss.) From the details of the spectra, and the relative height of the peaks and valleys, the gain can be determined.

Example: The laser above is 750 μm long and has facet reflectivity of 0.3 for both facets. For the peaks and valleys picture above and tabulated below, find the gain spectra over this wavelength range.

Valleys peaks			
Wavelength	Power (dBm)	Wavelength	Power (dBm)
1301.56	−61.22	1301.74	−57.87
1301.92	−61.93	1302.1	−58.3
1302.34	−61.73	1302.52	−57.94
1302.7	−61.85	1302.88	−57.47

The first thing to note is that the power is in dBm, which is a logarithmic unit. Power in mW is given by P(mW) = 10^P(dBm)/10. To take appropriate ratios for r_i, the power needs to be in linear units. To illustrate the calculation of just one point, the peak value at 1301.74 is 10^(−57.87/10), or 1.63 nW; the corresponding valley power is the average of −61.22 dBm (0.75 nW) and −61.93 dBm (0.64 nW), or 0.69 nW.

The ratio r_i is 1.63/0.69, or 2.36. The net gain g_{net} is $\frac{1}{750\times10^{-4}}\ln\left(\frac{2.36^{0.5}+1}{2.36^{0.5}-1}\right)+\frac{1}{2(750\times10^{-4})}\ln(0.3^2)=5\,\text{cm}^{-1}$.

Note that the first term is positive, representing gain; the second term is negative, representing mirror loss.

The rest of the points can be similarly calculated and give spectra as shown in Fig. 7.9. It is more interesting when plotted as complete spectra (across the whole range of available wavelengths), but a few points are all that is necessary to illustrate the technique.

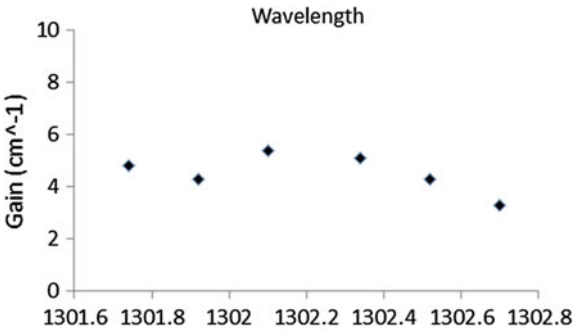


Fig. 7.9 Calculated gain spectra for a few points from the measured ratio of peaks to valleys

7.6 Lateral Modes in an Optical Cavity

The word 'mode' in an optical context is confusing because it means several things. It can mean 'wavelength', it can refer to the polarization state, or it can refer to the standing wave pattern inside an optical cavity in the propagation direction or the direction perpendicular to propagation. All these meanings are relevant to lasers, so let us clarify the particular modes we will be talking about getting into the details of each of them.

In Sect. 7.4, we discussed the longitudinal modes of a laser cavity. These are fairly easy to measure with an instrument like an optical spectrometer since each longitudinal mode corresponds to a slightly different wavelength.

But, in addition to the longitudinal modes, which identify the wavelengths in the cavity, there are lateral or spatial 'modes' that characterize the standing wave pattern of the light in the cavity transverse to the propagation direction. These are the same modes that characterize any waveguide. When we refer to a waveguide as 'single mode', this is the meaning of mode. Waveguides (including lasers) support many different wavelengths and are single mode in all of them.

In Sect. 7.3, we modeled a Fabry-Perot optical cavity as a single 1D slab of a single effective index. Here, we are going to look at the stacks of different materials that make up a laser section and see how they result in distinct modes each of which is characterized by a single effective mode index.

Figure 7.10 shows a simplified two-dimensional waveguide picture, with a region of higher index sandwiched by two regions of lower index. This is a slightly more realistic laser model than that in Fig. 7.1, since the quantum wells are of high

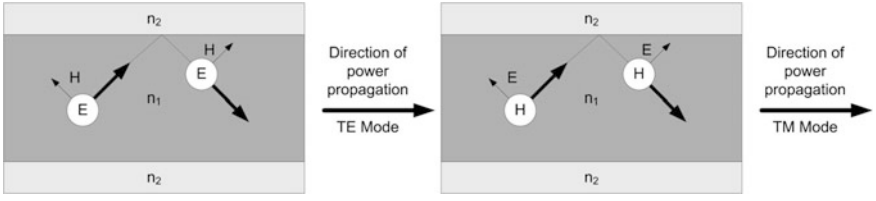


Fig. 7.10 Left, TE mode, and right, TM mode, propagating down a dielectric waveguide cavity

index than the cladding around them. This looks somewhat like a two-dimensional version of the Fabry-Perot waveguide; in that structure, the quantum wells in the middle serve as the waveguide as well as the means of carrier confinement. In this section, we will talk about the optical modes supported by the waveguide of Fig. 7.10.

Figure 7.10 shows a representation of the propagating modes in a waveguide. The direction of mode propagation is shown with a heavy arrow, and the orthogonal electric (E) and magnetic (H) field directions are indicated. The left figure shows the 'TE' mode, where the electric field is perpendicular to the direction of travel down the waveguide. Qualitatively, these optical modes are undergoing total internal reflection at the interface and bouncing back and forth between one side of the waveguide and the other. The quantitative details will be discussed shortly.

7.6.1 Importance of Lateral Modes in Real Lasers

Generally, for lasers used in communications, the waveguide structure is designed to realize a single transverse mode. Details of the design (like the thickness of the region around the cladding, or the etch depth of the ridge in a ridge waveguide device) are adjusted to achieve a device that is single mode. There are several reasons why this is important in semiconductor lasers.

First, as illustrated in Fig. 7.11, the mode shape also controls the far field of the device. Here the mode shape and far field pattern of a single-mode ridge waveguide device (right) and a broad area device (left) are compared. The far field pattern for a coherent light source is essentially the Fourier transform of the near-field pattern (which is the mode shape in the device.) The far field pattern of a single mode, ridge waveguide device is a fairly circular beam of modest, 30° divergence angle; the far field pattern of the broad area device is very elongated, with a few degree divergence in-plane and very high divergence out-of-plane. The pattern of optical power inside the cavity directly translates into the divergence pattern of light a few mm from the device. This is important because the ultimate objective of communications lasers is coupling into optical fiber, and for that purpose, a single-mode device is optimal.

Practically speaking, it is much easier to couple light between the relatively circular profile of a single-mode device and a fiber than the pattern of a broad area waveguide device.

The second reason it is important for a laser device to be single mode is that it is necessary for a device to be truly single wavelength. As we will learn in upcoming chapters, distributed feedback (DFB) devices make single-mode lasers using a periodic grating that reflects a single wavelength based on its effective index. Different lateral modes have different effective indexes, and therefore a multiple mode waveguide with a DFB grating could have more than one wavelength output.

A final practical comment is that, in reality, dielectric waveguides are only simple, first-order models for actual wave guiding of semiconductor lasers. The waveguide region of a laser is also the gain region, and so the refractive index has a

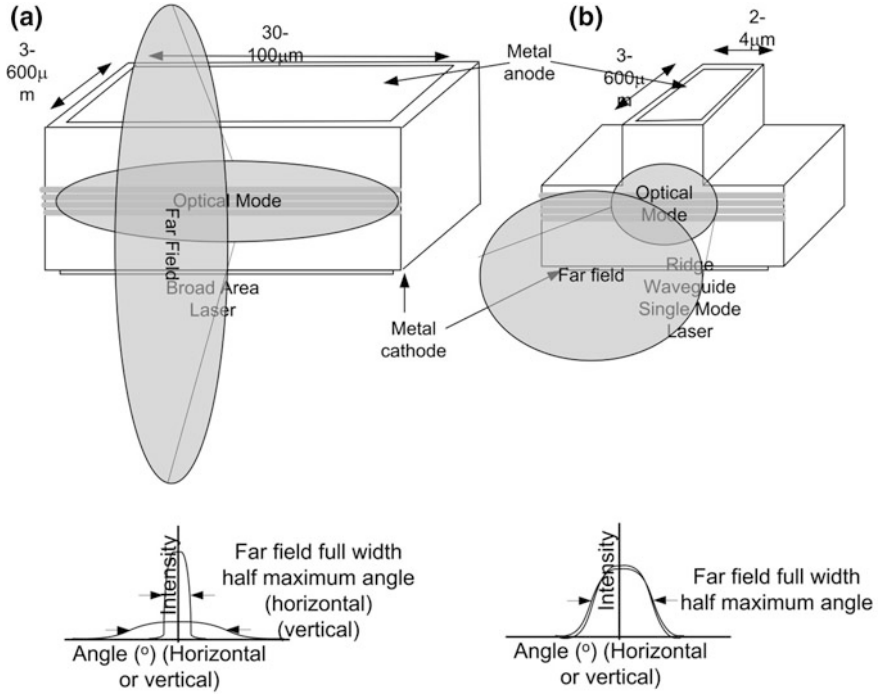


Fig. 7.11 Illustration of the importance of optical spatial mode by illustrating the dependence of far field on optical mode. **a** shows a broad area laser, several tens or hundreds of microns long; the top shows a schematic of the light exiting the laser, and the bottom shows a sketch of the intensity of the light vs divergence angle in the horizontal and vertical direction. A narrow horizontal stripe mode shape leads to a narrow vertical stripe far field. **b** shows a more circular single-mode device, with a nearly circular far field. Typical divergence angles of single-mode lasers are around 30°, though they can be engineered to be much lower

complex part associated with the gain (or, where there is no current, a loss component). The optical modes are said to be 'gain-guided' as well as index-guided, and really precise optical cutoff design is not required—this gain guiding tends to favor single-mode propagation. In practice, far fields and mode structure details calculated from index profiles can differ from the measurement of the fabricated device.

7.6.2 Total Internal Reflection

To get some insight into waveguide design, we are going to start with the idea of total internal reflection. As we hope the reader has previously encountered, when light is incident from a region of higher dielectric constant onto a region of lower dielectric constant, there is a critical angle. Light incident at angles above the

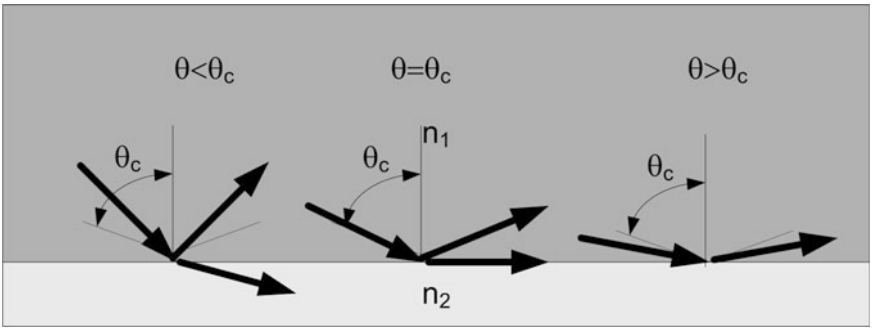


Fig. 7.12 Illustration of light inside a waveguide incident below, at, and above the critical angle, showing how a region of higher dielectric constant can act as a waveguide and conduct light down a channel

critical angle will glance off the side of the interface and experience total internal reflection. All of the optical power will be reflected at the incident angle. If the light is sandwiched between two such interfaces, the light will reflect back and forth between those interfaces and remain in the guiding region.

The formula for the critical angle θ_c is

$$\sin\theta_c = \frac{n_2}{n_1}. \tag{7.11}$$

Light incident above that angle θ_c will experience total internal reflection and remain within the cavity. Figure 7.12 illustrates what happens when light is incident on a dielectric interface at, below, and above the critical angle.

The picture shows a straightforward progression, in which the refraction away from the normal at the lower dielectric constant region goes from propagating into region 2 to propagating along the interface between the two regions, to propagation by total internal reflection inside region 1.

The above is a bit of a simplification. There is a little more subtlety associated with total internal reflection that explains some of its properties that we should at least qualitatively review.

First, it should be clear that the light has to interact a little with the low index region in order for it to 'see' it enough to be reflected by it. Light is a wave which occupies a length something like its wavelength. A more correct version of the total internal reflection picture shown at the right above might look like Fig. 7.13. The ray penetrates the material to a certain effective interaction length and then is reflected out. Because of this interaction length, a plane wave incident on a dielectric interface undergoes a phase change upon reflection. It can be pictured that the reflection at the point where the wave was incident actually comes from part of the plane wave incident slightly earlier, leading to what looks like an instantaneous phase shift.

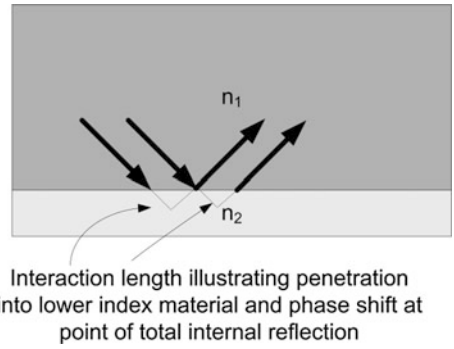

n1
n2
Interaction length illustrating penetration into lower index material and phase shift at point of total internal reflection

Fig. 7.13 A qualitative picture of the mechanism for phase shift at total internal reflection interface

Figure 7.13 implies that for a given ray, there should be a physical shift between its input and output. This effect actually happens with small, focused, light beams and is called the Goos-Hanchen effect. Though not particularly relevant in lasers, these sorts of effects are the reason that optics can be such a rich and fascinating subject although the basics of it have been known for centuries.

7.6.3 Transverse Electric and Transverse Magnetic Modes

In Fig. 7.10, modes with both transverse electric (TE) and transverse magnetic (TM) fields perpendicular to direction of propagation (hence, coming out of the page) are illustrated. In a waveguide, transverse is defined in terms of the guided waveguide direction, not in terms of the plane waves propagating inside the waveguide.

As a waveguide, a semiconductor laser will support both TE and TM modes, but in semiconductor quantum well lasers, the light emitted is predominantly TE polarized. The reason for that will be explored by Problem 7.3 and is based on the fact that the reflection coefficient at the facet differs for TE and TE modes. However, the result is that most laser light is inherently highly polarized.

For both TE and TM modes, only certain discrete angles can become guided modes which can travel down the waveguide. Just like light in an etalon has to undergo constructive interference in order for the etalon to support a particular wavelength, light in a waveguide has to undergo constructive interference for a particular 'mode' (which corresponds to a particular incident angle) to exist. In an etalon analysis, usually the variable is wavelength, and transmission is plotted as a function of wavelength; in a waveguide analysis, typically the wavelength is fixed, but nature chooses the angle at which it propagates. The reason for it is also the same; assuming the plane wave in the cavity originates from all the points on the bottom edge, if the round trip weren't an integral number of wavelengths, destructive interference would eventually cancel that optical wave.

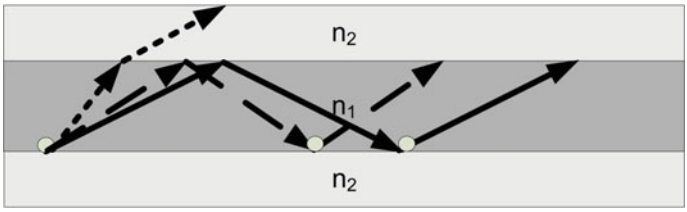


Fig. 7.14 An example of two allowed propagating modes. The white dots are points with a 2π phase difference. Other possible modes, represented by the more dotted lines, have an incident angle below the critical angle for that particular dielectric interface and so are not allowed

As is illustrated in Fig. 7.13, in addition to the phase change due to propagation, there is also a phase change at total internal reflection. Both of these phase changes must be taken into account when determining the allowed waveguide modes.

Figure 7.14 shows two allowed modes using arrows. The definition of an allowed mode is that the net phase difference between the two equivalent points be an integral multiple of 2π.

If the waveguide is a higher index region sandwiched by two identical lower index regions, there is always at least one very shallow angle in which this condition is satisfied. Depending on the index difference and thickness, there may be other angles which also fulfill this condition. Eventually, the incident angle will exceed the critical angle and the necessity of total internal reflection will not be met. The quantitative aspect of determining the allowed modes will be discussed in the next section.

7.6.4 Quantitative Analysis of the Waveguide Modes

In this section, we will go through calculation of guided guides for some simple waveguide structures. The purpose is to give a more intuitive picture of what a mode is, not to present the best calculation techniques. Nowadays, software is usually used to obtain modes for lasers or most complicated wave guiding structures. The reader is invited to look at other books (e.g., Haus[2]) for examples of waveguide solutions by other methods.

The qualitative picture now should be clear. Transverse electric or transverse magnetic (TE or TM) modes can both simultaneously propagate in a higher index medium sandwiched by two lower index mediums. For a symmetric medium (with the same index cladding region on both sides), there is always at least one allowed propagation angle and one guided mode. As the index contrast gets higher, the critical angle gets higher and the number of modes increases. A thicker higher index region also increases the potential number of modes.

[2]H. Haus, *Waves and Fields in Optoelectronics*, Prentice Hall, 1984.

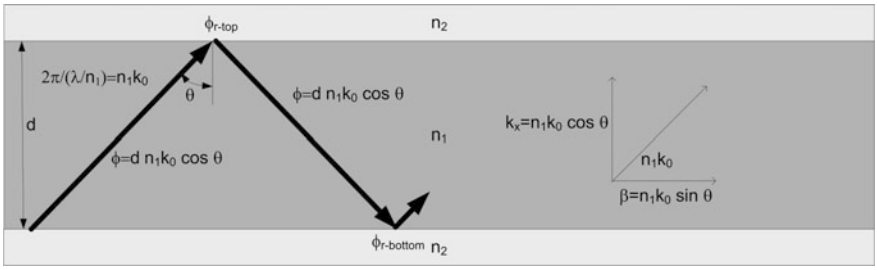


Fig. 7.15 A waveguide illustrating the phase change of a propagating mode at reflection and due to the propagation length. The propagation constants in the forward and up-and-down direction are identified in terms of the fundamental propagation constant $2\pi/(\lambda/n_1)$

Figure 7.15 identifies the angles and propagation constants in various directions, and the phase changes at reflection. The top and bottom slabs are considered to be infinitely thick. The propagation constant k_0 of light in free space is

$$k_0 = \frac{2\pi}{\lambda} \tag{7.12}$$

On examination of this figure, let us write down the mathematical statement that the net phase change between equivalent parts of the wave, the far left and the middle, should be a multiple of 2π. The relevant quantities are defined in the figure.

$$2\phi + \phi_{r-\text{top}} + \phi_{r-\text{bottom}} = 2dn_1k_0\cos\theta + \phi_{r-\text{top}} + \phi_{r-\text{bottom}} = 2m\pi \tag{7.13}$$

where the ϕ terms are the phase changes due to reflection (defined below). Put in a different way, the round trip from bottom to top should be an integral number of wavelengths, even though the light is propagating mostly forward. For light which is mostly forward, the phase change is given by k_x (the k vector in the x-direction) multiplied by the distance, which is $n_1k_0\cos\theta$. Conventionally, the propagation constant in the forward direction is called β, and it is equal to $n_1k_0\sin\theta$.

The phase change on total internal reflection is

$$\phi_{\text{TE}} = -2\tan^{-1}\left(\frac{\sqrt{n_1^2\sin^2\theta - n_2^2}}{n_1\cos\theta}\right) \tag{7.14}$$

for TE waves, and

$$\phi_{\text{TM}} = -2\tan^{-1}\left(\frac{n_1\sqrt{n_1^2\sin^2\theta - n_2^2}}{n_2^2\cos\theta}\right) \tag{7.15}$$

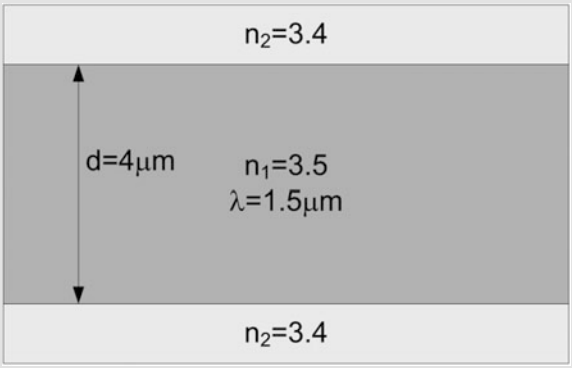
for TM waves.

The effective index n_{eff} which identifies the mode is given by

$$n_{\text{eff}} = n_1 \sin \theta_p \tag{7.16}$$

where θ_p now means that we have identified a particular discrete propagating angle as labeled in Fig. 7.15. Let us illustrated this process of analyzing propagation waveguide modes with an example, and then discuss more qualitatively what design variables are adjusted to tailor a single-mode waveguide.

Example: Find the number of TE modes, and the effective index of all the TE modes, supported by the waveguide pictured.



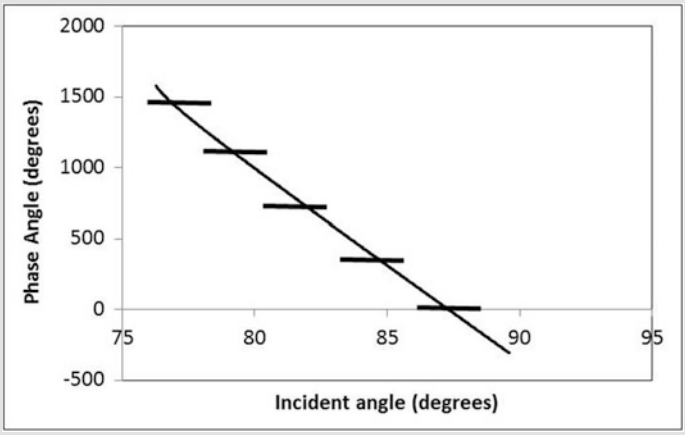
Solution: The equations are formulated in terms of k (the propagation vector) and θ (the incident angle from high-index region to the low index region, measured from the normal). The propagation vector $k = (3.5)2\pi/(1.5 \times 10^{-6}) = 14.66 \times 10^6/\text{m}$. Equation 7.13, written with known quantities and an angle θ, is

$$2(4 \times 10^{-6})3.5(4.83 \times 10^6)\cos\theta - 4\tan^{-1}\left(\frac{\sqrt{3.5^2 \sin^2\theta - 3.4^2}}{(3.5)\cos\theta}\right) = 2m\pi$$

Finding the allowed modes in the waveguide corresponds to finding the allowed values of θ in the equation above. The equation is a transcendental equation. There is no analytic solution, and the effective way to solve it is to plot the left side versus θ and pick out the values of θ for which the equation is true.

The range of theta is set by the expression under the square root sine. When $\theta = \sin^{-1}(3.4/3.5) = 76.3°$ the angle becomes greater than the critical angle, and the mode is no longer reflected by total internal reflection. Only angles

between 90 and 76.3 have to be considered. The graph below plots the left side of the expression above, with lines indicating the multiples of 360° points (including 0).



The line has the following phase angles at the following incident angle. At each angle, the propagation constant β is given by $k\sin\theta$, and the effective index n_{eff} is given by $\beta n_1/k$.

Phase angle θ	Incident angle	β (/m)	n_{eff}
0	87.3°	14644477	3.496114
360	84.7°	14598072	3.485036
720	82.0°	14518073	3.465938
1080	79.4°	14410570	3.440273
1440	77.0°	14284995	3.410294

There are five modes in the waveguide as listed above.

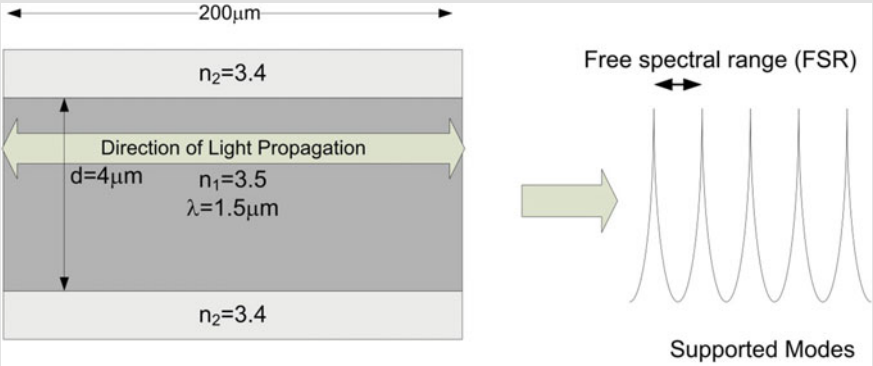
It is important to look at the example above and try to get some qualitative insight. First, notice how the effective index ranges from 3.49 to 3.41 (between the value of 3.5, the value of the high-index guiding layer, and 3.4, the lower index, cladding layer). At the shallow angle of 87.3°, the optical mode is traveling mostly straight down the guiding layer, and effectively 'seeing' mostly the index of the guiding layer. At the steeper angles, with the mode bouncing more often between the two sides, it sees more of the cladding. The effective index is closer to the cladding. It is the effective index, not the material layer index that governs the properties of the waveguide and is used, for example, in the expression above for cavity finesse (Eq. 7.2, and other expressions with n).

Every high-index layer surrounded by symmetric low index cladding has guided modes—at least one each TE and TM mode. As the layer gets thicker, or the index contrast gets higher, the number of guided modes in a structure increases.

For lasers, generally thicker more confining waveguides are better, since better confinement to the active region leads to lower thresholds and better overall properties. However, as the waveguide gets thicker and higher confining, it gets more multimode. As with many things in lasers, designing the waveguide is a tradeoff. The goal is usually to get the thickest single-mode waveguide possible.

Finally, let us do a final example to connect the one-dimensional etalon in Sect. 7.4.2, with this two-dimensional waveguide here.

`Example: Find the free spectral range of the lowest order mode of the simple dielectric waveguide structure below.`



`Solution: The formula for free spectral range is given in Eq. 7.6, and the only question is what index to use. The appropriate index is the mode index for the structure above. As the geometry is the same as the previous example, the index of the lowest order mode is 3.496114, and the free spectral range is then`

$$\Delta\lambda = \frac{\lambda^2}{2Ln} = \frac{1.5^2}{2(200)3.496114} = 1.61\,\text{nm}.$$

The one-dimensional structure of Sect. 7.4.2 could be considered a model of a more realistic, two-dimensional waveguide shown here. The mode indexes determined by the waveguide govern the optical output.

There are other, equivalent formulations for determined the discrete modes of a slab waveguide that are involved matching boundary conditions at the boundary, which is perhaps more flexible in the case of more than three layers. In practice,

much of this optical modeling is done in software, and this simple three-layer method illustrates clearly the origin of discrete modes without being too computational.

7.7 Two-Dimensional Waveguide Design

We are going to extend Sect. 7.4 into another dimension. Instead of looking at light confined in the *y*-direction while it travels in the *z*-direction, we will now look at light confined in *y*- and *x*-direction, while it travels back and forth in the *z*-direction. This is an accurate picture of what happens in a laser cavity.

7.7.1 Confinement in Two Dimensions

A typical laser waveguide, like the ridge waveguide structure whose cross-section is shown in Fig. 7.11, left, (and in the example problem below) is actually a two-dimensional confining structure. One can think of the light being confined in the *y*-direction by the higher index of the active region compared to the cladding region, like a typical slab waveguide. How is it confined in the *x*-direction?

The answer is subtle and best seen by imaging the optical mode as a diffuse blob that is centered on the confining slab but leaks out to the cladding and the ridge above. When this optical mode overlaps with the ridge, it sees a higher average index than to the left and right, where the mode overlaps more with the air. This index difference between the effective index of the center, where the ridge is, and the effective index on the sides, where the top layer is removed and the optical mode sees only air, forms the optical confinement in the x-direction.

In ridge waveguide structures like this, typically the index difference in the *x*-direction is much less than the index difference in the *y*-direction. In such circumstances, numbers for the optical mode as a whole can be more easily obtained by the effective index method, which we will illustrate (again, largely by example) in the sections below.

7.7.2 Effective Index Method

Below we are going to illustrate a more manual method for solving simple indexes for two-dimensional confinement regions. (In reality, these calculations grow extraordinarily complicated with multiple layers and real shapes actually seen in lasers, and so real calculations are usually done using programs, such as RSOFT or Lumerical. This example will illustrate at least how the geometry and index contrast determine whether a waveguide is single mode or not).

Table 7.2 Analyzing waveguides using the effective index method

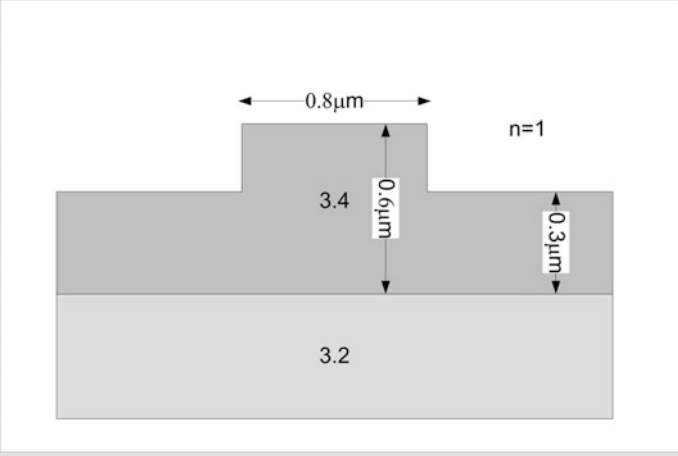
Steps for analyzing simple ridge waveguide-type structures using the effective index method
1. Break the waveguide up into two regions (inner and outer) and solve for the effective modes of each of those regions, the chosen polarity
2. Make a slab waveguide using those effective indexes as the core and cladding index
3. Find the effective index of that simple structure, which is approximately the effective index of the 2D waveguide

For pedagogical reasons, let us model the typical semiconductor waveguide as shown below in the upcoming example. A region of about 3.4 effective index is clad by air (on top) and a semiconductor substrate (3.2) on the bottom. In a ridge waveguide geometry, the region around the central region is etched to provide confinement in the *x*-direction.

The basic process for the effective index method is shown in Table 7.2.

This method works well if the confinement in one direction (typically in the *y*-direction) is much stronger than in the *x*-direction.

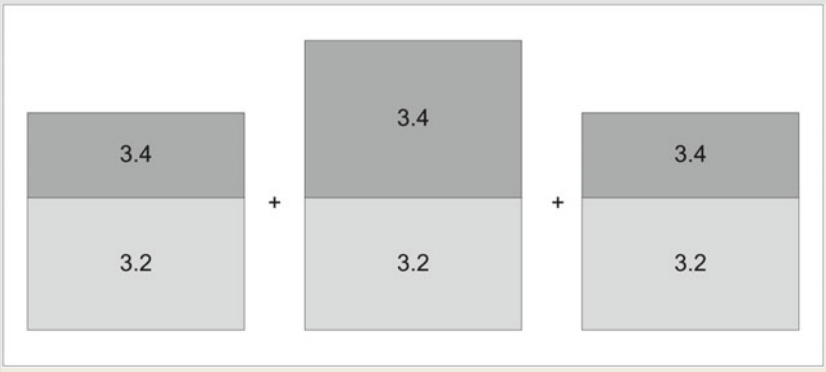
Example: Find the effective index (or indexes) of the TE modes of light at a wavelength of 1.3 μm confined in the ridge waveguide structure below



Solution: First, we break the structure into three separate structures, as shown. Equation 7.13 applies to each structure, but of course, the phase change (and critical angle) at the top and bottom interface is different. Equation 7.13, for example, written for the middle slab, would be:

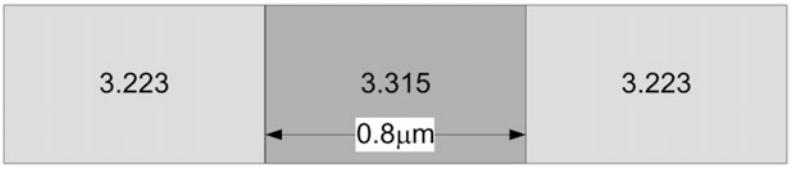
$$2(0.6\times10^{-6})(3.4)(4.83\times10^{6})\cos\theta - 2\tan^{-1}\left(\frac{\sqrt{3.5^2\sin^2\theta-3.4^2}}{(3.5)\cos\theta}\right) - 2\tan^{-1}\left(\frac{\sqrt{3.5^2\sin^2\theta-1^2}}{(3.5)\cos\theta}\right) = 2m\pi$$

Solving Eq. 7.13 for each of the slab indexes results in the following values for TE effective modes in each slab.



n=3.223 n=3.315 n=3.223

Finally, the waveguide in the *x*-direction looks approximately like the structure below.



Solving this structure in the *x*-direction leads to the following effective index: $n = 3.281$. Since all of the structures in the example were single mode, the final result is also single mode. If the width were 1 μm instead of 0.8 μm, the final structure would have had two modes, 3.289 and 3.223. Since the objective is to have the widest structure that is still single mode, a target ridge width should be between 1 μm and 0.8 μm.

7.7.3 Waveguide Design Targets for Lasers

Now that we know how to analyze index structures for wave guiding modes, let us discuss what the optimal waveguide for a laser should look like. For the sake of discussion, let us draw a picture of a simple ridge waveguide, let the ridge width vary, and see what happens to the effective index n and the mode shape.

As shown below in Fig. 7.16, with a very narrow ridge, the effective index is close to the cladding index. This implies that the optical mode is very large and 'sees' a lot of the cladding. (Qualitatively, the effective index n_{eff} is some kind of weighed average of the indexes that the mode shape covers.) For lasers, the optical mode should be confined to be in the gain region (indicated by the dark region under the ridge) where the quantum wells are and where the injected current produces gain. As the ridge gets wider, the effective index sees more of the region under the ridge and gets slightly higher, and the optical mode is more confined to the region under the ridge. Finally, as the ridge gets wider yet a second mode appears. This second mode has a two-peak standing wave pattern in the ridge.

For lasers, the best waveguide is the most confining, single-mode device. High confinement to the region under the active region means net high optical gains and lower threshold currents. Multimode devices as discussed can have worse coupling to optical fiber and not be single wavelength.

As we close this section, and chapter, let us make a final comment. While discussion of the mathematics of how to calculate optical modes gives insight into what influences the optical mode, usually, real-mode solutions for complex structures are done with numerical methods on software such as Lumerical or RSOFT. The analytic analysis of a waveguide with many, many parts is very difficult.

7.8 Summary and Learning Points

In this chapter, we discuss the influence of the cavity on the light. A typical laser structure with two reflecting facets sandwiching an active region acts as an etalon, and only allows certain wavelengths within the cavity. This allowed wavelengths form the set of longitudinal modes.

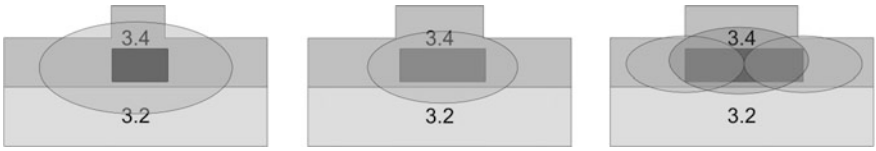


Fig. 7.16 Illustration of mode shape evolution versus ridge width in a simple example. The less confined modes (left) have bigger modes and worse confinement to the active region (indicated by the dark rectangle). In the middle just before cutoff, the optical mode is most confined to the active region. Finally, on the right, a second mode appears, characterized by two peaks. The ideal design target for lasers is just before the single-mode cutoff, illustrated in the middle

In addition, the details of the wave guiding structure including the index contrasts and dimensions control the spatial modes of the devices. These modes can influence the wavelengths supported by the cavity, and control the coupling into and out of optical fiber.

With the tools of this chapter, waveguides can be designed to support only a single spatial mode. With that, truly single-wavelength devices, using, for example, distributed feedback structures, can be fabricated.

A. In an optical cavity defined by two mirrored surfaces, only certain wavelengths are supported due to constructive/destructive interference between the facets.
B. Each supported wavelength in a cavity must have an integral number of round trips between the two facets.
C. A Fabry-Perot laser cavity has a regular spacing of modes determined by the length of the cavity.
D. The number of wavelengths is given by the cavity length and the mode index; spacing between wavelengths depends on the group index.
E. Each supported lasing wavelength is identified as a longitudinal mode in a Fabry-Perot laser.
F. The number of lasing modes is determined by the gain bandwidth and the mode spacing.
G. A laser cavity is also a waveguide composed of a higher index region sandwiched by lower index regions
H. The laser waveguide supports one or more transverse/spatial/lateral modes.
I. These modes are found for a system with one-dimensional confinement by finding the discrete angles at which light reflected back and forth undergoes constructive interference from the top to the bottom.
J. The specific angles each correspond to a different mode.
K. The effective index method can be used for systems with two-dimensional confinement in which the index contrast in one direction is much less than in the other direction (as in typical ridge waveguide lasers).
L. Although mathematically TE and TM modes are equally supported in a waveguide, real semiconductor laser emit predominantly TE light because the facet reflectivity is slightly higher (and the distributed facet loss slightly lower) for TE light.
M. Laser waveguides should be designed to be just before the cutoff for single-mode waveguides. They should have the highest possible effective index before the waveguide becomes multimode.
N. Real-mode solutions for complex structures are usually done with numerical methods on software such as Lumerical or RSOFT.
O. Lasers are gain-guided as well as index-guided. Often the details of the effective index and far field differ from those calculated using index guiding alone.

7.9 Questions

Q7.1. What is an etalon?
Q7.2. What modes are supported in an etalon?
Q7.3. What is a difference between an etalon and a Fabry-Perot laser cavity?
Q7.4. What is the expression for the spacing between allowed modes in a cavity?
Q7.5. What is the expression for the number of wavelengths in a cavity?
Q7.6. What is the difference between the group index and the index? Why does the group index determine the mode spacing?
Q7.7. What is the condition for a lateral mode?
Q7.8. Does every high-index structure sandwiched by low index regions support at least one mode?
Q7.9. Is it possible for an index waveguide to support a TE mode but not a TM mode, or a TM mode but not a TE mode?

7.10 Problems

P7.1. Derive Eq. 7.6 and then Eq. 7.8 for free spectral range, appropriate for vacuum and semiconductor etalons, respectively.
P7.2. Write Eq. 7.6 in terms of optical frequency, v, rather than wavelength.
P7.3. A InP-based laser emitting at λ = 1550 nm has a 300 μm cavity length, a group refractive index n = 3.4, and refractive index of 3.2. The width of the gain region above threshold is 30 nm.

(a) What is the mode spacing, in

(i) nm?
ii) GHz?

(b) How many modes are excited in the cavity?
(c) What is the typical number of wavelengths in a round trip in the cavity?

P7.4. Semiconductor lasers typically emit strongly polarized light. If the facet reflectivity for an incident angle of θ (from the perpendicular) is given by

$$R_{\mathrm{TE}} = \frac{n_1 \cos\theta_i - n\cos\theta_t}{n_1 \cos\theta_i + n\cos\theta_t}$$

for TE polarized modes, and

$$R_{\mathrm{TM}} = \frac{n_1 \cos\theta_{ti} - n\cos\theta_i}{n_1 \cos\theta t + n\cos\theta_i}$$

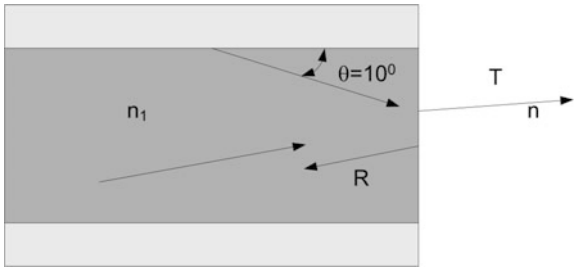


Fig. P7.17 Laser modes incident on the facet in a semiconductor waveguide

for TM polarized modes, calculate the reflection coefficient for TE and TM modes, and the associated distributed facet loss, for the modes pictured in Fig. P7.17 (let $n_1 = 3.5$ and $n = 1$). What polarization do semiconductor lasers emit? (Hint: consider the distributed facet loss for each polarization).

P7.5. The ring laser pictured in Fig. P7.18 is a triangular waveguide fabricated on a piece of quantum well semiconductor material. Two of the facets are etched at an angle for total internal reflection, so that the entire light wave is reflected. The other angle ingle is made more abrupt so that the incidence angle is below that needed for total internal reflection. The light goes around the ring which serves as the cavity, and the arrows show (one) direction of light circulating in the ring.

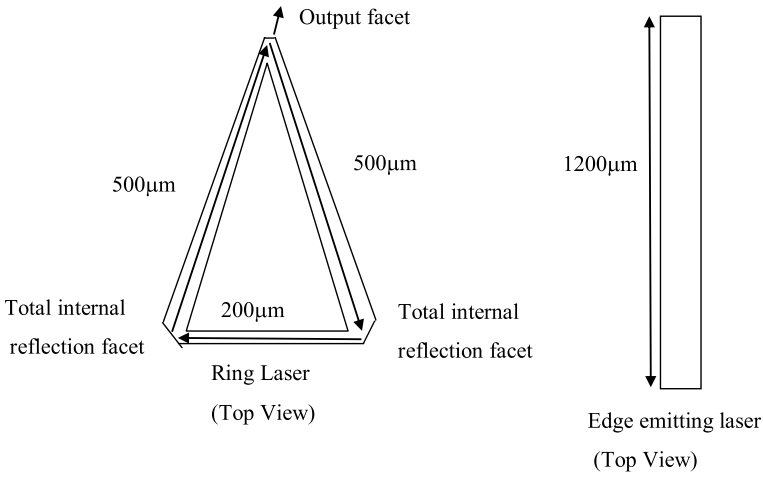


Fig. P7.18 A triangular ring laser (left) and a conventional edge-emitting laser (right)

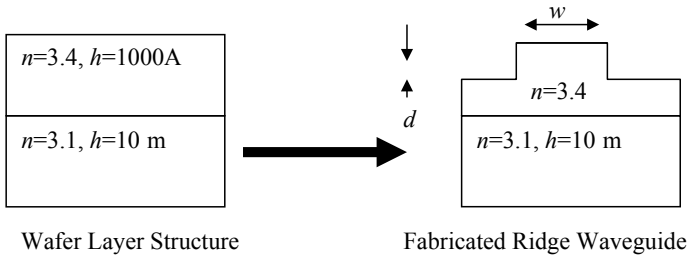


Fig. P7.19 Waveguide design problem

The group index is 3.5, the mode index is 3.2, and the lasing wavelength is 1.3 μm.

(a) If the long legs of the triangle are (as pictured) 500 μm, and the short leg is 200 μm, what is the expected mode spacing in the device?
(b) Which device would have a greater threshold current (the ring laser or the edge emitting) given that they are the same 'size' and facet reflectivity on output facets (and, briefly, why)?

P7.6. Assume a waveguide is formed by a layer of 3.5 index core, 2 μm thick, surrounded by cladding with a refractive index of 3.2 (as in the example of Sect. 7.6.4, with a different thickness). Find the number of TM modes, and the incident angle and effective index associated with each mode.

P7.7. A very simple optical model of a waveguide structure is given in Fig. P7.19, consisting of a higher index layer on top of a lower index layer (sandwiched by air on top). Determine an etch depth and rib width to make this structure a single-mode 'rib' waveguide as shown. (Note: there are many possible answers!).

P7.8. Look back at Problem 6.10, where the question was what doping would be necessary to reduce the resistance of the top contact to 5Ω. Another thing that a designer could do to decrease resistance is to increase the top contact width.

(a) What width for the ridge would be necessary to reduce the resistance to 5Ω?
(b) What optical problems could that possibly cause in laser operation?

Laser Modulation

8

He said to his friend, "If the British march
By land or sea from the town to-night,
Hang a lantern aloft in the belfry arch
Of the North Church tower as a signal light–
One if by land, and two if by sea;
And I on the opposite shore will be"
—Henry Wadsworth Longfellow, *Paul Revere's Ride*

Abstract

In this chapter, the use of lasers for direct modulation transmission at high speeds is discussed. The laser properties that limit the high speed transmission and the ultimate transmission speed achievable are analyzed.

8.1 Introduction: Digital and Analog Optical Transmission

Semiconductor lasers in optical communications are largely used as digital modulated light sources. Just as Paul Revere doubled the light in Longfellow's famous poem, lasers are switched from low light levels to high light levels to communicate digital zeros or ones in an optical fiber. The data on the fiber is encoded in little pulses of light which then travel at the speed of light down the flexible optical fiber waveguide. Because so much information can be transmitted on the fiber, we (the end user) have as much bandwidth as we are willing to pay for (with more available all the time).

As discussed in the previous chapters, the optical power output from a laser is proportional to the current injected into the laser. In the simplest digital amplitude modulation scheme, high level light pulses represent '1's and low level light pulses represent '0's. In a direct modulation scheme, these '1's and '0's are generated by rapidly switching the current injected into the laser between two different levels. In

© Springer Nature Switzerland AG 2020

D. J. Klotzkin, *Introduction to Semiconductor Lasers for Optical Communications*,
https://doi.org/10.1007/978-3-030-24501-6_8

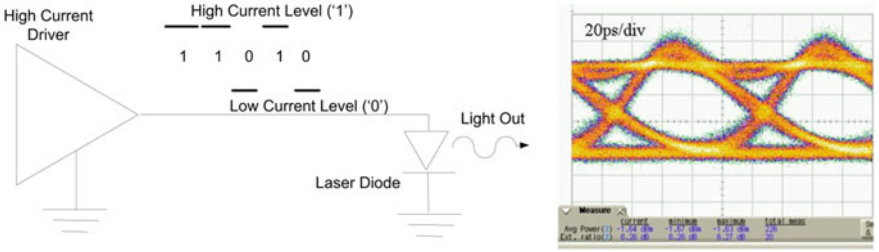


Fig. 8.1 Left, a simplified directly modulated laser diode circuit. Right, a typical eye pattern showing changing light levels in response to a random pattern of 1's and 0's. The region in between illustrates the digital current data (clean transitions between the 1 and 0 current levels) versus the output light data

this chapter, we discuss the limits of the speed with which we can directly modulate lasers. Chapters 11 and 12 talk about laser transmission systems on a higher level, including externally modulated devices and an overview of transmitters and receivers.

To illustrate what we mean by 'modulation,' Fig. 8.1 shows the laser output in the form of an *eye pattern* (which is the conventional way that large signal digital optical modulation schemes are evaluated). An eye pattern shows many bits overlaid on each other, in which each bit starts at the same point on the trace. A desirable eye pattern has a clean transition between high and low, looking (in fact) like the square current pulse that drives the laser. The very typical laser output shown in Fig. 8.1 looks nothing like that. It has significant overshoot, much slower rise and fall times and is delayed from the input current pulse. These properties result from semiconductor laser characteristics and fundamentally affect how a semiconductor laser can be used for direct modulation.

We hope this brief introduction to eye patterns was, at least, eye-opening.

Alternatives to direct laser modulation include external modulation, in which a laser is used to generate the source light and another modulation method is used to change the light amplitude.

Before we discuss the fundamental limits of digital transmission, let us look at the requirements on optical digital transmitters. This will tell us what the semiconductor lasers *have* to do before we focus on what they can do.

8.2 Specifications for Digital Transmission

It is worthwhile to discuss how digital transmission is specified and to connect the transmitter specification to the laser bias conditions and coupling efficiency from the laser to the output fiber. To avoid the situation in which vendors test their products to many slightly different standards, the industry has tried to provide common standards for optoelectronic components. Table 8.1 shows a bit of the

Table 8.1 Typical specification for a directly modulated optical transmitter

Parameter	Minimum	Maximum	Typical
Wavelength at 25 °C (nm)	1290 nm	1330 nm	1310 nm
I_{th}(25 °C) (mA)	5	20	10
SMSR (dB)	35 dB	–	40 dB
Coupled slope efficiency (W/A)	0.1 W/A	–	0.2 W/A
Launch power (dBm)	−8.5 dBm	0.5 dBm	−3 dBm
Extinction ratio (dB)	3.5 dB	–	–

specification for a laser component designed to be used as part of the IEEE 802.3 compliant transponder.

Power in laser specifications is often given in dBm (decibel mW) as given below.

$$P(\text{dBm}) = 10\log\left(\frac{P(\text{mW})}{1\,\text{mW}}\right) \tag{8.1}$$

For example, 0 dBm is 1 mW, 10 dBm is 10 mW, and so on. The *extinction ratio* is the ratio of the power at the '1' level (P_{on}) to the power at the '0' level (P_{off}). This is usually given in dB:

$$\text{ER}(\text{dB}) = 10\log\left(\frac{P_{on}}{P_{off}}\right) \tag{8.2}$$

The specification on extinction ratio implies a specification on laser speed. When the extinction ratio is given, it means the transmitter should pass a mask test (as will be described below) at that given extinction power. Qualitatively, the eye should look open at that speed and bias conditions, with an acceptable amount of overshoot and a blank area in the middle so the receiver can decide if it is receiving a zero or a one.

For 1550 nm directly modulated devices, another specification on laser high speed modulation is dispersion penalty, which will be discussed in Chap. 10.

The launch power, LP, means the average fiber-coupled power, given by

$$\text{LP} = 10\log\left(\frac{P_{on} + P_{off}}{2\,\text{mW}}\right) \tag{8.3}$$

in dBm. It differs from the laser power because the laser (in whatever packaged form it is being sold) does not couple all of the light out into a fiber. Only a certain fraction of light emitted from the front facet of the laser (typically around 50%, though it can be higher) is translated into useful transmittable light.

Given the value of extinction ratio, launch power, and laser characteristics, the necessary bias conditions can be determined. An example of the calculated bias conditions I_{high} and I_{low} is given below.

Example: A typical laser has a threshold current of 10 mA and a coupled slope efficiency of 0.15 W/A into the fiber. For typical transmission conditions (LP = −1 dBm, extinction ratio of 4 dB for a 10 Gb/s device), calculate the low and high current levels.

Solution: From the expression

$$-1 = 10\log_{10}(\mathrm{LP(mW)}/1\ \mathrm{mW}),$$

the launch power is calculated to be 0.8 mW. The power of 0.8 mW is an average current of

$$0.8\,\mathrm{mW}/0.15\,\mathrm{W/A} = 5.33\,\mathrm{mA} = \left(I_{\mathrm{high}} + I_{\mathrm{low}}\right)/2.$$

above threshold. By the extinction ratio given

$$4 = 10\log_{10}\left(P_{\mathrm{high}}/P_{\mathrm{low}}\right) = 10\log_{10}\left(0.15I_{\mathrm{high}}/0.015I_{\mathrm{low}}\right),$$

the ratio of I_{high}/I_{low} is 2.5. From the average current expression,
$(I_{high} + I_{low})/2 = (2.5I_{low} + I_{low})/2 = 5.33$ mA or $I_{low} = 3$ mA (above threshold) and $I_{high} = 7.6$ mA (above threshold).

In this chapter, we are going to focus on the factors that limit directly modulated laser speed and how to get a fast device. We start by talking about small signal modulation (which is useful in its own right and often a good figure of merit for large signal communication) and then connect it to large signal properties. Then we talk about other limits to high speed transmission, including fundamental laser characteristics and more parasitic limitations.

8.3 Small Signal Laser Modulation

In some applications, the laser is used directly in an analog small signal transmission mode. For lasers used to optically transmit cable TV signals (CATV lasers), the channel information is actually encoded into analog modulation of the laser output. Though the small signal characteristics are directly relevant here, usually the modulation frequency is very low compared to the laser capabilities.

Typically, the small signal characteristics are used to describe the laser speed metrics, but the device is used digitally.

We first describe a small signal measurement and then discuss its application, first to light-emitting diodes (LEDs) and then to lasers.

8.3.1 Measurement of Small Signal Modulation

Before discussing the theory of small signal modulation, let us illustrate the modulation measurement, so the reader can have a good idea of the properties being measured and relate to the upcoming mathematics.

When we talk about modulation bandwidth of lasers and LEDs, what we mean is the frequency response of the quantity $\Delta L/\Delta I$, where L is the light output and I is the input current. In these measurements, the device (laser or LED) is typically DC-biased to some level, and an additional small signal amount of current is superimposed on this DC bias. The amplitude of the small signal light is then measured and plotted as a function of frequency, and the point where the amplitude falls 3 dB below the DC or low frequency response is called the device bandwidth. The measurement and frequency response are illustrated in Fig. 8.2.

These measurements are much easier to describe and quantify than large signal measurements. It is not clear immediately how to put a number to how good an eye pattern is, but it is quite straightforward to name a device bandwidth under a certain DC bias condition.

These small signal measurements are important measurements for lasers for several reasons. First, they give direct information about the physics of the device, including information about the optical differential gain that cannot be obtained directly. They also serve as a good proxy for large signal measurements: devices with good (high) bandwidths give good eye patterns.

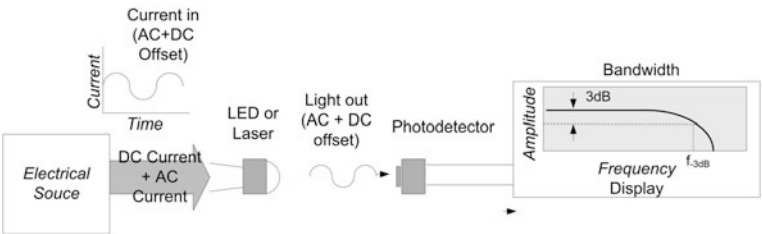


Fig. 8.2 Illustration of a modulation measurement for an optical device (laser or LED). The device is DC-biased, and a small signal is superimposed on top of it. The small signal amplitude of the light is plotted against frequency to give the device bandwidth. Sometimes, the source and receiver are in the same box, called a network analyzer. The bandwidth is the point where the response falls to 3 dB below its low frequency level

8.3.2 Small Signal Modulation of LEDs

To enter into this subject of large signal laser modulation, let us begin by small signal modulation of light-emitting diodes. This will serve to give a more intuitive picture of what determines modulation bandwidth of these devices and introduce the small signal rate equation model that we will use to model these phenomena.

The simplest meaningful model includes electron and hole current injection into the active region and radiative recombination in the active region. Figure 8.3 illustrates the processes.

Figure 8.3 neglects carrier transport and leakage through the active region, but captures the important details. The important concept is that the carrier population in the active region is only increased by increased current and only decreased by radiative recombination. When a certain current level is applied to the device, a certain DC level of carriers is established in the active region. The carrier population in the active region can only increase through current injection and only decrease through recombination, which has a time constant, τ_r, associated with it. Inherently and intuitively, the bandwidth should be limited by that recombination time constant.

A rate equation that describes the process is given in Eq. 8.4,

$$\frac{dn}{dt} = \frac{I}{qV} - \frac{n}{\tau}. \tag{8.4}$$

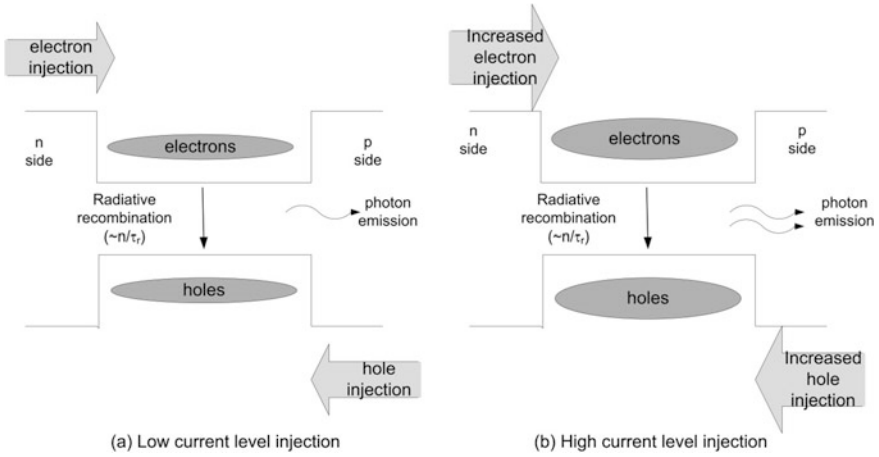


Fig. 8.3 Modulation of LEDs. Current is injected into the active region, where it recombines radiatively and emits light. Modulation speed is limited because once in the active region, the current density reduces only with the ~ns timescale associated with radiative recombination. The figure shows **a** modest carrier population density and light output with low level current injection, and **b** increased carrier population density and light output with higher level current injection

In the equation, n is the carrier density in the active region, I is the injected current, V is the volume of the active region, q is the fundamental unit of charge, and τ is the carrier lifetime. That carrier lifetime in this simple model represents the amount of time it takes before a carrier radiatively recombines into a photon.

The first term in Eq. 8.4, I/qV, represents injected current; the second term, n/τ, represents carriers which recombine after a time τ and emit a photon and hence is proportional to the photon emission rate S_{emission} out,

$$S_{\text{emission}} = n/\tau_{\text{r}} \tag{8.5}$$

in which τ_{r} is the *radiative lifetime*. The radiative lifetime is the lifetime of the carriers due to the process of radiative recombination only. Total carrier lifetime τ is the carrier lifetime due to both radiative (τ_{r}) and non-radiative (τ_{nr}) processes. If the processes are all independent, the total lifetime is given by Matthiessen's Rule as

$$\frac{1}{\tau} = \frac{1}{\tau_{\text{r}}} + \frac{1}{\tau_{\text{nr}}}. \tag{8.6}$$

The radiative efficiency η_{r}, which is the fraction of injected carriers which are emitted as photons, is given as,

$$\eta_{\text{r}} = \frac{\frac{1}{\tau_{\text{r}}}}{\frac{1}{\tau_{\text{nr}}} + \frac{1}{\tau_{\text{r}}}}. \tag{8.7}$$

Problem 8.1 will explore the implications of these different times. For now, let us note that the internal quantum efficiency of a good laser can be >90%, and in both laser and LED material, radiative recombination dominates.

To model a small signal measurement, both I and n are given a DC and an AC component (at a frequency ω), as shown in Eq. 8.8.

$$\begin{aligned} I &= I_{\text{DC}} + I_{\text{AC}} \exp(jwt) \\ n &= n_{\text{DC}} + n_{\text{AC}} \exp(jwt). \end{aligned} \tag{8.8}$$

Let us substitute these expressions for I and n into the simple rate equation of Eq. 8.4 to obtain

$$\frac{\mathrm{d}n_{\text{DC}} + \mathrm{d}n_{\text{AC}}}{\mathrm{d}t} = \frac{I_{\text{DC}} + I_{\text{AC}}}{qV} - \frac{\mathrm{d}n_{\text{DC}} + \mathrm{d}n_{\text{AC}}}{\tau}, \tag{8.9}$$

which breaks up into two simple equations. One of them,

$$0 = \frac{I_{\text{DC}}}{qV} - \frac{n_{\text{DC}}}{\tau}, \tag{8.10}$$

sets the DC carrier level in the diode as a function of injected bias,

$$n_{\mathrm{DC}} = \frac{I_{\mathrm{DC}}\tau}{qV}. \tag{8.11}$$

The second AC equation,

$$n_{\mathrm{AC}} j\omega \exp(j\omega t) = \frac{I_{\mathrm{AC}} \exp(j\omega t)}{qV} - \frac{n_{\mathrm{AC}} \exp(j\omega t)}{\tau} \tag{8.12}$$

can be rewritten, by canceling the common exponential term and rearranging, as

$$\frac{n_{\mathrm{AC}}}{\frac{I_{\mathrm{AC}}}{qV}} = \frac{1}{1 + j\omega\tau} \tag{8.13}$$

$$\left| \frac{n_{\mathrm{AC}}}{\frac{I_{\mathrm{AC}}}{qV}} \right| = \frac{1}{\sqrt{1 + \omega^2\tau^2}}. \tag{8.14}$$

The only step necessary in recognizing this as the modulation bandwidth of an LED is to realize that the light out is proportional to the current density n.

Notice that this is an experimental prescription for measuring the carrier lifetime. It is exquisitely difficult to observe carriers directly, but it is perfectly straightforward to measure the 3-dB bandwidth of a device. From that bandwidth (assuming that the measurement is unobstructed by parasitics, and there are no other meaningful recombination terms), the carrier lifetime can be extracted.

There can, of course, be a phase offset between n and I (represented by a complex n_{ac}) but it is irrelevant to measuring the bandwidth.

8.3.3 Rate Equations for Lasers, Revisited

What we wanted to show in the previous discussion is that LED modulation was fundamentally limited by the carrier lifetime in the active region because their fundamental emission mechanism is spontaneous emission from carrier recombination. As the carrier lifetime is of the order of ns, the lifetime is limited to ranges typically <1 GHz.

Lasers, however, emit light by stimulated emission. The stimulated emission lifetime is much shorter than spontaneous emission, as the carrier recombination is controlled by the changing photon density. The expectation is that therefore laser modulation will be fundamentally different and faster.

As with LEDs, let us start with a rate equation, with appropriate small signal terms inserted. The appropriate rate equations (from Chap. 5) are repeated below.

$$\begin{aligned}\frac{\mathrm{d}n}{\mathrm{d}t} &= \frac{I}{qV} - \frac{n}{\tau} - G(n,S)S \\ \frac{\mathrm{d}S}{\mathrm{d}t} &= S\left(G(n,S) - \frac{1}{\tau_\mathrm{p}}\right) + \frac{\beta n}{\tau_\mathrm{r}}\end{aligned} \tag{8.15}$$

Most terms are defined as before: I is the current injected, V is the active region volume, τ and τ_r are the total recombination time and radiative recombination time, respectively; τ_p is the photon lifetime and β is the fraction of carriers coupled into the lasing mode. The final term is generally important only in kick-starting the laser process; once the optical gain becomes non-negligible, it is the spontaneous emission photons that are amplified to generate the lasing photons.

The one change we make is a redefinition of gain function $G(n, S)$, which is now given as a function of both carrier density n and photon density S. In Chap. 5, we were looking at the DC steady state value of the gain and for that purpose, a DC value sufficed. Here, when we want to include time dependence, we need to use a more sophisticated model, which includes the carrier density and the photon density.

$$G(n,S) = \frac{\mathrm{d}g}{\mathrm{d}n}(n - n_\mathrm{tr})(1 - \varepsilon S). \tag{8.16}$$

This model incorporates two important physical factors. The differential gain $\mathrm{d}g/\mathrm{d}n$ is an important metric for high speed laser performance. What it represents is the change in gain with increase in carrier density. Though the DC gain is clamped at threshold, modulating the laser involves changing the current in resulting in a change in light level out. This $\mathrm{d}g/\mathrm{d}n$ parameter measures how quickly this happens, and thus how fast a device can be modulated.

This model also assumes that the model is strictly linear all the way from transparency through and around lasing. This is a simplification, but usually perfectly applicable.

The gain function also includes the 'gain compression' factor ε. This factor models the fact that as the current into the laser increases (above threshold), the net AC gain that the light in the cavity experiences decreases. For example, at low photon/current levels into a laser, a temporary increase in carriers may increase the output (temporarily, until the steady state DC situation is restored) by a hypothetical 10%; the same increase in carrier density at high photon/current levels may only increase the output by 5%. This excess carrier density can be directly created by modulation of the input current or by optical pumping.

It is safe to say that the mechanisms for gain compression are not fully understood and vary depending on the details of the laser structure. Two of the common mechanisms for gain compression are illustrated in Fig. 8.4. The first is called spectral hole burning, in which the carrier distribution becomes nonlinear as the photon density gets higher and depletes carriers at the lasing wavelength. The second is called spatial hole burning, in which the higher photon density at certain locations (at the facets, in a Fabry-Perot laser or anywhere, in a distributed feedback

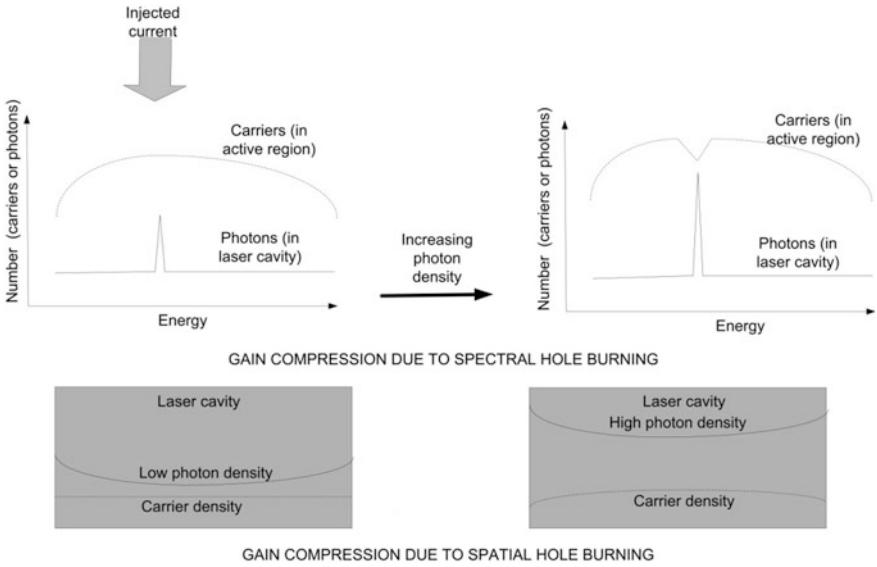


Fig. 8.4 Mechanism for gain compression. The top illustrates spectral hole burning, in which the current density becomes non-equilibrium as the light intensity increases, leading to a reduction in effective gain at the lasing wavelength; the bottom illustrates spatial hole burning, where locations with high photon density have non-uniform carrier density

laser) depletes the carriers at those locations and reduces the net gain. Whatever the mechanism, this gain compression at higher currents and photon densities damps out the modulation response.

A word about the units: in the rate equations, gain is in units of 1/s, and differential gain is in units of 1/s-cm^3. When gain is calculated using band structures, conventionally it is in units of 1/cm, and differential gain is in units of cm^2 (change in gain in 1/cm divided by carrier density in 1/cm^3). It can easily be converted from one to another by multiplying by the group velocity.

$$\begin{aligned} G(n,s)\left[s^{-1}\right]/v_g[\mathrm{cm/s}] &= G(n,s)\left[\mathrm{cm}^{-1}\right] \\ \frac{\mathrm{d}g}{\mathrm{d}n}\left[\mathrm{s}^{-1}\mathrm{cm}^3\right]/v_g[\mathrm{cm/s}] &= \frac{\mathrm{d}g}{\mathrm{d}n}\left[\mathrm{cm}^2\right] \end{aligned} \tag{8.17}$$

In the rate equation context of this chapter, both these numbers should be understood to be in s^{-1} units. And note how useful unit analysis can be helpful in navigating complicated equations!

8.3.4 Derivation of Small Signal Homogenous Laser Response

To begin talking about the dynamic response of lasers, let us first solve for the small signal homogenous laser response. From the rate equations, we write the appropriate, small signal differential equations for n_{ac} and s_{ac}, where the 'ac' subscripts indicate deviations from the DC solution. Here, we will follow Bhattacharya's[1] treatment, slightly simplified

$$S = S_{DC} + s_{ac}$$

$$n = n_{DC} + n_{ac}. \tag{8.18}$$

The variable n_{dc} is n_{th}, which is usually a few times the transparency current density n_{tr} for a given structure. At this point, the math gets complicated, so let us describe what we are going to do first before we go ahead and do it.

(i) Substitute the expressions in Eq. 8.15 into the rate equation, Eq. 8.12. The resulting equation will have first order terms containing single terms n_{ac} or s_{ac}, zeroth order terms which contain neither and second order terms which contain products of n_{ac} and s_{ac}.
(ii) Ignore the second order terms (considering them generally small compared to the first order terms) and the zeroth order terms (since those are exactly the DC rate equations!).
(iii) Finally, we are going to write differential equation for ds_{ac}/dt and dn_{ac}/dt. This equation is appropriate for when the steady state conditions for n or s are perturbed and will describe how the laser evolves back to the steady state. It will give some insight into the dynamics of the laser.

As a real example, let us take the rate equation for n and apply these steps.

$$\frac{dn_{DC} + dn_{ac}}{dt} = \frac{I}{qV} - \frac{n_{DC} + n_{ac}}{\tau} - \frac{dg}{dn}(n_{DC} + n_{ac} - n_{tr})(1 - \varepsilon(S_{AC} + S_{DC}))(S_{AC} + S_{DC}) \tag{8.19}$$

The following results can be carried through including the gain compression ε, but they are much more complicated. To give the following expressions a bit more intuition, the ε term will be henceforth set to zero and we will leave out the spontaneous emission term from the photon rate equation. We will also set the drive term (I) to zero to find the homogenous solutions.

Setting ε equal to zero and keeping only the first order, small signal terms on both sides give

[1]Pallab Bhattacharya, *Semiconductor Optoelectronic Devices*, 2nd edition, Prentice Hall.

$$\frac{dn_{ac}}{dt} = -\frac{n_{ac}}{\tau} + S_{AC}\frac{dg}{dn}(n_{dc} - n_{tr}) - n_{ac}\frac{dg}{dn}S_{DC} \tag{8.20}$$

And

$$\frac{dS_{ac}}{dt} = -S_{ac}\left(\frac{dg}{dn}(n_{dc} - n_{tr}) - n_{ac}\frac{dg}{dn}S_{DC} + \frac{1}{\tau_p}\right). \tag{8.21}$$

These two equations are a set of coupled, linear differential equations; ds_{ac}/dt and dn_{ac}/dt depend on s_{ac} and n_{ac}. The reader is reminded that the DC gain is clamped at threshold and does not vary. The ac value of n and s, and the total gain, do vary.

The equations can be combined into a single second order differential equation by differentiating one of them (say, the equation for ds_{ac}/dt), and substituting for dn_{ac}/dt in the first equation then an expression can be obtained that contains only the first and second derivatives of s. We leave the details of that operation to the curious reader. The homogenous solutions are of the usual exponential form

$$n(t) = \exp(-\Omega t)\exp(j\omega_r t), \tag{8.22}$$

which looks like a decaying sinusoid. By using the DC expressions (for example,

$$\frac{1}{\tau_p} = \frac{dg}{dn}(n_{th} - n_{tr}), \tag{8.23}$$

which can be obtained from setting the rate equation for s equal to zero above threshold, as done in Chap. 5), fairly simple expressions for Ω and ω_r can be written. The time constant of the decay, Ω, can be written as

$$\Omega = \frac{1}{2\tau}\left(\frac{i}{i_{th} - i_{tr}}\right), \tag{8.24}$$

where

$$i_{th} = \frac{n_{th}q}{\tau} \tag{8.25}$$

and

$$i_{tr} = \frac{n_{tr}q}{\tau}. \tag{8.26}$$

The resonance frequency is then equal to

$$\omega_r = \left[\frac{1}{\tau\tau_p}\left(\frac{i}{i_{th} - i_{tr}}\right) - \Omega^2\right]^{1/2}, \tag{8.27}$$

which since τ_p (the photon lifetime of ps) $\ll \tau$ (the carrier lifetime of ns), is approximately

$$\omega_r = \left[\frac{1}{\tau\tau_p}\left(\frac{i}{i_{th} - i_{tr}}\right)\right]^{1/2}, \tag{8.28}$$

The relaxation frequency is a geometric average of the photon and carrier lifetime and increases as the square root of the bias current. Both of these will ultimately affect the design of lasers and their chosen operating points for high speed operation. Typically, small devices, with short photon lifetimes, are faster, and we will see that higher speed devices are specified to a lower extinction ratio and higher currents on the low end.

8.3.5 Small Signal Laser Homogenous Response

Equation 8.21 spells out the form of the natural response of a laser when there are small variations from the DC parameters. For example, if, in an active laser, a pulse of light injected a small number of excess carriers above the DC value, that equation would describe how the carriers (and the light) decayed down to their equilibrium values.

To illustrate this in operation, see Fig. 8.5. This figure shows how a laser responds when current is suddenly applied. The figure does *not* show the small signal solution; it is a full numerical solution of the rate equation response of Eq. 8.12, essentially the large signal response. However, the tail end of the response when the current and light are converging toward their steady state values is characteristic of the small signal solution that we determined above. The form of the response shows what the natural response looks like.

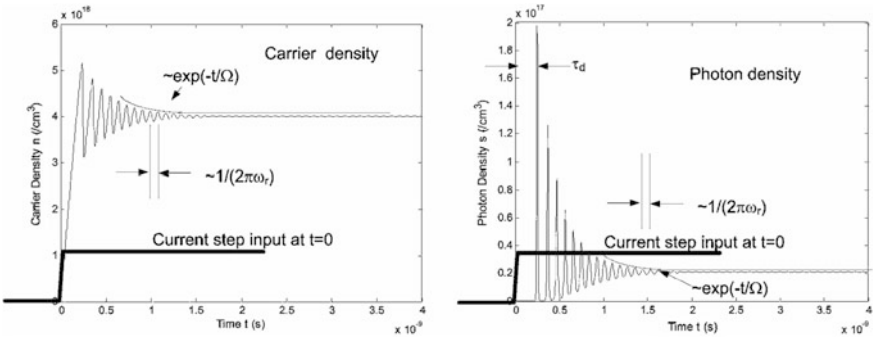


Fig. 8.5 An illustration of the nonlinear solution of the rate equation illustrating what happens when the current to a laser is abruptly turned on

In this calculation, at time $t = 0$, the current in goes from 0 amps to some nonzero value, above threshold. The figure on the left illustrates what happens to the carrier density in the active region. After the current starts, carriers start to accumulate in the active region, until carrier density approaches the threshold carrier density. In steady state, excess current above threshold turns into photons, not carriers; however, several nanoseconds elapse before the population of carriers and photons equilibrates. During that time, the population of carriers and photons oscillates as it decays t its equilibrium value.

The 'why' of it requires some explanation. Until the carriers reach threshold, there are very few photons created by spontaneous emission. Hence, there is a delay between when the current input starts and when the light output begins. Above threshold, the net positive gain results in a sudden increase in photons which results in a depletion of carriers. The population of photons oscillates at the same frequency as the carriers, and they both gradually decay to their equilibrium value.

For both photons and carriers, as the difference from equilibrium value gets small, the response looks like the small signal response. The decay time $1/\Omega$ and the resonance frequency ω_r can be identified by the spacing between oscillations and the falloff of the peaks, as shown.

This is the *fundamental* reason that bit patterns of high speed lasers have the sort of overshoots that are shown in Figs. 8.1 and 8.11. These oscillations are inherent to directly modulated lasers. Typically, the receiver is low-pass-filtered to improve the response and reduce the impact of these typically high frequency oscillations.

8.4 Laser AC Current Modulation

With what we know about the natural response of the laser system, we can start to discuss the modulation response of a laser. The small signal modulation response is the small signal change in output light L (or photon density S) due to a small signal change in input current, I, plotted versus frequency. The measurement is precisely the same as illustrated in Fig. 8.2.

8.4.1 Outline of the Derivation

The outline of the derivation of the laser modulation response equation is given here, though we spare the reader the grittiest of details.

First, to determine an expression for laser modulation response, we start by letting the I in the rate equation have both an AC and DC component, as shown below.

$$I = I_{DC} + i_{ac} \exp(jwt) \tag{8.29}$$

The AC amplitude i_{ac} is at a frequency ω, which models modulating the device at a frequency ω. This time-varying input leads of course to a time-variation in optical output and in carrier density.

If the AC term is small compared to the DC term, a small signal approximation is appropriate and the output terms (n and s) should now have a form

$$n = N_{DC} + n_{ac} \exp(j\omega t), \tag{8.30}$$

and

$$s = S_{DC} + s_{ac} \exp(j\omega t) \tag{8.31}$$

also with AC and DC components.

From here, the process is similar to that illustrated in determining the laser natural small signal response in Sect. 8.3.4. The rate equations are expanded and the first order terms (including just $\exp(j\omega t)$) are retained, leading to a first order rate equation in the small signal quantities s_{ac}, n_{ac}, and i_{ac}. The response s_{ac} for a fixed magnitude i_{ac} can be found as a function of the modulation frequency ω. That expression is the modulation response, which can be related to the experimental measurement illustrated in Fig. 8.2. The experimental measurement and the equation are developed in the next few sections. In the derivation, the inclusion of the gain compression term ε is necessary to model laser behavior accurately.

8.4.2 Laser Modulation Measurement and Equation

Let us start by illustrating in Fig. 8.6 a typical small signal laser modulation measurement, as pictured in Fig. 8.2, followed by the equation that it should match. The measurements are at room temperature at different currents and illustrate the typical shape of the response. The dots illustrate the measured response, and the curves are 'best fits' to the theoretical expression which will be discussed.

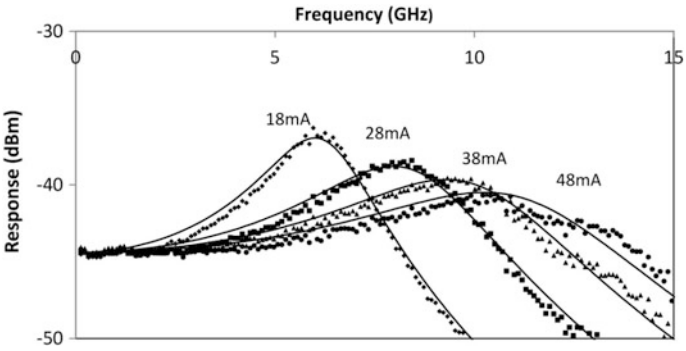


Fig. 8.6 Bandwidth data (points) and best fit curve (line) to Eq. 8.32

Qualitatively, the responses for most semiconductor lasers are similar. As the DC current into the device increases, the resonance peak moves out in frequency as the height of the peak gets lower.

Both of these effects are accurately predicted from the small signal model of the laser rate equations,

$$M(f) \prec \frac{1}{\left(f^2 - f_r^2\right) + j\frac{\gamma}{2\pi}f} \frac{1}{1 + j2\pi f\tau_c} \tag{8.32}$$

To more easily match the output of a standard network analyzer, this equation is given in terms of frequency f, rather than angular frequency ω, which is $2\pi f$. The parasitic term τ_c comes from a more complete rate equation model which includes transport and parasitics (see Problem 8.5): it will be discussed below. The damping factor term, γ, is defined in Eq. 8.34. Most of the complex laser behavior under small signal modulation is contained in this fairly simple equation (and in the two other equations that we will discuss in this section). The modulation response looks like a second order function with a resonance peak (representing the fundamental laser response at f_r) along with a first order additional falloff, representing parasitic terms. As the laser current increases, the resonance peak f_r increases also, according to

$$\begin{aligned} f_r &= \frac{1}{2\pi}\left[\frac{1}{\tau\tau_p}\left(\frac{i}{i_{th} - i_{tr}}\right)\right]^{1/2} = \frac{1}{2\pi}\sqrt{\frac{S_{DC}}{\tau_p}\left(\frac{dg}{dn} + \frac{\varepsilon}{\tau}\right)} = \frac{1}{2\pi}\sqrt{\frac{v_g\frac{dg}{dn}(I - I_{th})\eta_i}{qV}} \\ &= D\sqrt{I - I_{th}} \end{aligned} \tag{8.33}$$

Equation 8.33 shows the dependence of resonance frequency on current or photon density written in several common ways. Fundamentally, increasing the photon density increases the resonance frequency. However, it is difficult to measure the photon density in the cavity directly (but see Problem 8.2!), so the second expression includes the conversion of light to current, and charge to carriers, with the internal conversion efficiency η_i (the fraction of injected carriers which are converted to photons) and to the electron charge q.

Accurately knowing the photon density S_{DC} or the carrier active volume V is quite difficult. What is often measured is simply the relationship between injected current and measured resonance frequency, which theoretically and experimentally follows the quadratic form given. The symbol D (the laser *D-factor*, in GHz/mA$^{1/2}$) is then a metric for laser performance.

The damping, γ, which describes how the peak flattens out, is given by

$$\gamma = \frac{1}{\tau} + Kf^2, \tag{8.34}$$

where K is the damping factor, given by

$$K = (2\pi)^2 \left(\tau_p + \frac{\in}{\frac{dg}{dn}} \right) \tag{8.35}$$

Physically, this damping term arises because the modulation is limited by photon lifetime (which is significant at high frequencies and high photon densities, even though it is typically much smaller than carrier lifetime) and by gain compression, which are the two terms in Eq. 8.35. The easiest one to picture is photon lifetime: just as carrier lifetime fundamentally limits processes driven by carrier population (such as LED light emission), photon lifetime in lasers fundamentally limits modulation bandwidth. Gain compression also acts to reduce the bandwidth. As current is injected, the gain both increases (because of dg/dn) and decreases (because the photon density increases and the gain is reduced due to gain compression). Thus, the effective differential gain becomes less at high bias currents.

Finally, the final term in the expression $(1/(1 + j2\pi f\tau_c))$ is a model for both the parasitic RC time constant and for carrier transport into the active region of the laser diode. The first part of the expression models the behavior of the laser active region. To completely model the effects, the frequency limits of injecting carriers into the active region also have to be included.

The physical picture origin of τ_c is shown in Fig. 8.7.

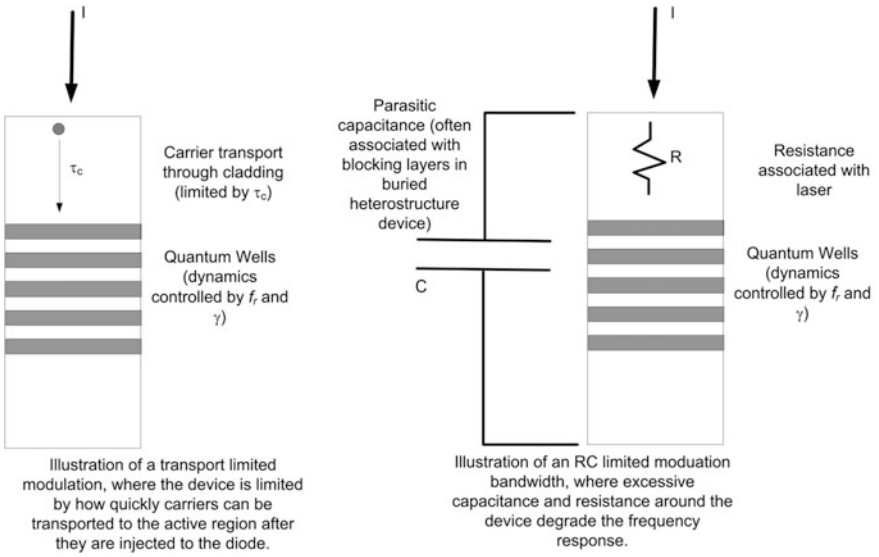

I
τc
Carrier transport through cladding (limited by τc)
Quantum Wells (dynamics controlled by fr and γ)
Illustration of a transport limited modulation, where the device is limited by how quickly carriers can be transported to the active region after they are injected to the diode.
I
Parasitic capacitance (often associated with blocking layers in buried heterostructure device)
C
R
Resistance associated with laser
Quantum Wells (dynamics controlled by fr and γ)
Illustration of an RC limited moduation bandwidth, where excessive capacitance and resistance around the device degrade the frequency response.

Fig. 8.7 Illustration of transport limited bandwidth (left) and RC limited bandwidth (right). In both cases, the modulation response is degraded due to factors external to the laser active region

Transport is the easiest to imagine. The carrier, injected into the high resistance, low-doped regions of the diode, typically takes a few ps to make its way to the active region. If the cladding is exceptionally thick, the diffusion across it can take more than a few ps and so directly affects the modulation bandwidth directly.

Excessive RC transport constants can give rise to the same behavior. Typical laser diodes have a few ohms of resistance associated with them (about 8–12 Ω for 300 μm devices) due to current flow through the moderately doped p-contact and cladding region. If the diode has excessive capacitance associated with it as well, the modulation response sees what looks to be a single-pole, low-pass RC-filter. This impacts the modulation bandwidth in the same way.

This capacitance can come from capacitance associated with the blocking layers (in buried-heterostructure devices) or from the metallization layers or from the junction. Resistance and capacitance can typically be adjusted by adjusting those external factors (doping or metallization patterns) while keeping the same laser active region.

Both of these effects are included in the laser modulation by including an additional rate equation with the two shown in Eq. 8.12. This equation represents carriers injected into the cladding directly by the current and then transported to the active region in a characteristic time τ. (The reader is asked to write down the appropriate rate equation in Problem 8.5.)

8.4.3 Analysis of Laser Modulation Response

After the data is acquired, typically the data is analyzed. The method for analyzing the data will be illustrated in the example below.

Example: From the data in Fig. 8.6 (for which the best fit is shown tabulated), determine the D- and the K-factor and estimate the differential gain and gain compression for a device. The device is a Fabry-Perot device with uncoated facets and a 200 μm long cavity, a 2 μm wide ridge and a total active region (including quantum wells and barriers) of 130 nm. The absorption loss in the material (which has been previously measured) is 20 cm^{-1}. The effective mode index is 3.2.

Solution: The first step is to fit the data obtained to the theoretical curve. When that is done, using with the data above, the following fit parameters (or ones close to them) are obtained:

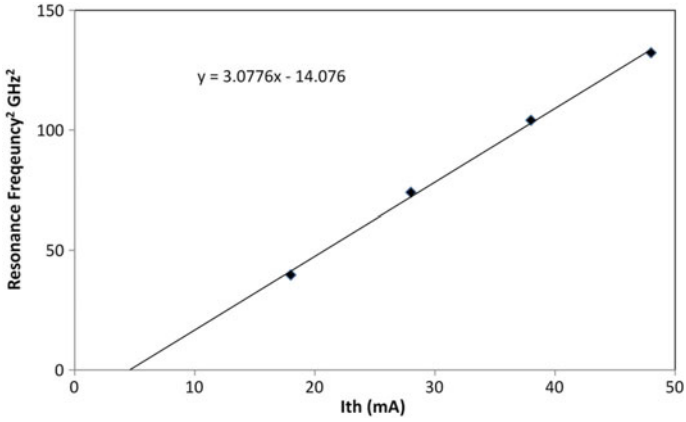


Fig. P8.8 The data for resonance frequency^2 versus current, showing an extrapolated threshold current of around 5 mA and a slope, in Ghz^2/mA of 3.07 and a *D-factor* of 1.75 GHz/mA$^{0.5}$

Current, I	f_r (Ghz)	γ (1/s)	τ_c (ps)
18	6.3	16	10
28	8.6	26	10
38	10.2	33	10
48	11.5	44	10

Based on Eq. 8.33, the square of the resonance frequency is plotted versus the injected current in Fig. P8.8. To find differential gain, the form of Eq. 8.33 is used.

$$\frac{f_r}{\sqrt{(I - I_{th})}} = 1.7 \times 10^9 = \sqrt{\frac{v_g \frac{dg}{dn} \eta_i}{qV}}$$

From the dimensions, the active volume is 5.5×10^{-11} cm^3 and the group velocity (c/n) is 9.4×10^9 cm/s. Hence, the differential gain is 1.0×10^{-15} cm^2. (The units for differential gain look unusual; remember that it is changed in gain, in 1/cm, divided by change in carrier density, in 1/cm^3).

According to Eq. 8.33, the x-intercept is the threshold current. For this particular device, the threshold current is about 5 mA. This threshold (determined from measurements of the modulation response only) typically agrees well with the threshold obtained from L-I measurements.

To find gain compression, first, the measured γ versus f_r^2 is plotted. From Eq. 8.34, the slope is K and the y-intercept is $1/\tau$. The points are shown in Fig. P8.9.

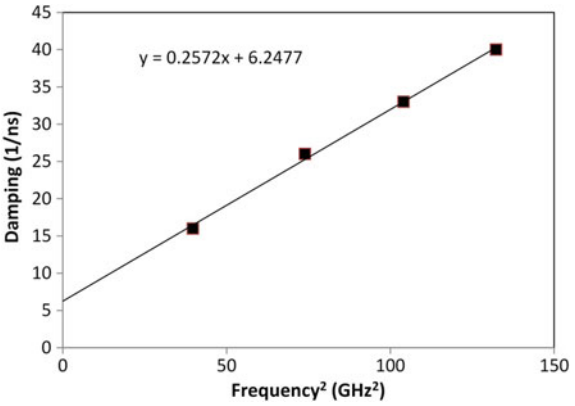


Fig. P8.9 Damping factor, in 1/ns, versus resonance frequency squared. The slope gives the *K-factor* (in ns), and the intercept gives one of the carrier lifetimes

Here, the *K-factor* is 0.25 ns and the carrier lifetime is about 0.2 ns. To get a number for photon lifetime which also appears in Eq. 8.15, we use the DC rate equation,

$$g_{\text{modal}} = 1/\tau_{\text{p}}.$$

The modal gain is given by the sum of optical loss + material loss or

$$g_{\text{modal}} = \frac{1}{2L}\ln\left(\frac{1}{R^2}\right) + \alpha = \frac{1}{2(200 \times 10^{-4})}\ln\left(\frac{1}{0.3^2}\right) + 20 = 80\,\text{cm}^{-1}$$

The photon lifetime $\tau_{\text{p}} = 1/(80\ v_{\text{g}}) = 1.3$ ps. Plugging all this information into Eq. 8.35 gives $\varepsilon = 4.9 \times 10^{-18}$ cm^{-3}. These units indicate the photon density at which the gain is meaningfully compressed. At a photon density of $>10^{17}$ cm^3, the gain will be reduced by 5% or more, according to Eq. 8.16.

This example hopefully illustrates the typical process of looking at a laser response and analyzing its dynamics. It also illustrates how one can use measurements to get to fundamental material quantities. In this case, straightforward measurements on bandwidth give differential gain and gain compression, which are intrinsic properties of the active region. The method, used here and everywhere in science and engineering, is to relate measurement quantities to material properties using an appropriate model. Terms like dg/dn that are indirectly measured from an

analysis of laser bandwidth, for example, can be directly tied to the theory considering the band structure of the device.

The appropriateness of the model can be empirically judged by the goodness of fit between the data and the model (shown in Fig. 8.6). In this case, the fit is reasonably good. If the fit is generally poor (for this model or for anything), it is usually wise to reexamine the model. In general, this modulation model here (with three fitting parameters per curve, f_r, γ and τ_c) is a good model to measured laser response.

8.4.4 Demonstration of the Effects of τ_c

In analysis of this device, we obtained a τ_c value of about 10 ps, roughly independent of bias current. This τ_c represents the RC time constant associated with the device (as well as transport time associated with carrier injection from the highly conductive contact layers to the active region).

Typical lasers have a resistance associated with them on the order of 5–10 Ω (sometimes more), so this level τ_c represents an associated capacitance of about 1 pF. This is a reasonable value considering typical geometric capacitances associated with laser metal pads or the reverse-biased capacitance associated with blocking structures in buried heterostructure regions.

The influence of this capacitance term on the laser performance is often underestimated and sometimes even omitted from analysis of laser response. Figure 8.10 shows the results of an experiment in which the capacitance was intentionally varied by varying the size of the metal contact pad on the laser surface. Depending on the structure, this pad typically has some capacitance associated with it equal to $\varepsilon A/d$, where d is the distance to the doped chip surface and A is the metal pad area.

As can be seen, the laser modulation response differed enormously as this parasitic capacitance was intentionally varied. While generally a high bandwidth is preferred (which implies a minimal capacitance), sometimes a flat response is desirable. In that case, the capacitance can be optimized to improve the response as desired.

8.5 Limits to Laser Bandwidth

Laser bandwidths are limited by both intrinsic factors, contained in the modulation equation and other factors. The two factors which are included in the modulation equation are the *K-factor* limit, and the transport and capacitance limit.

The number K encapsulates how quickly the peak flattens out as it moves out in frequency. The units of K are time (typically, ns). This damping by itself can limit the laser bandwidth. This limit is appropriately called the damping limited bandwidth $BW_{damping}$ and is given by

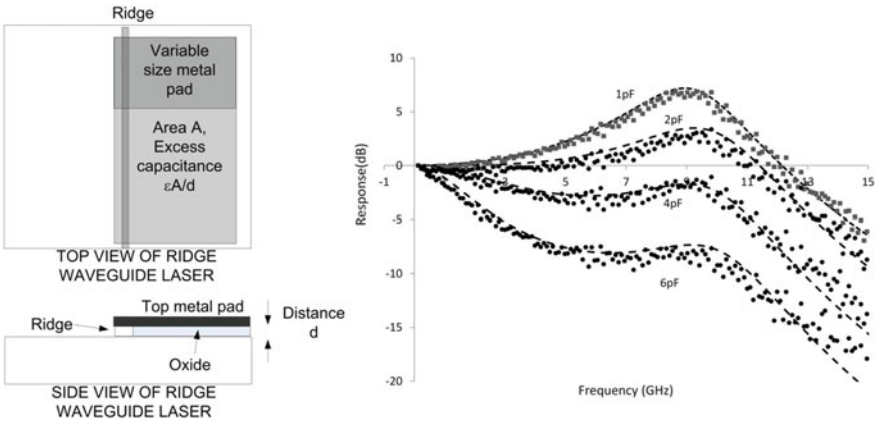


Fig. 8.10 Left, a description of an experiment in which many identical lasers were fabricated with differences in the size of the compliant metal pad, which typically sits on an oxide on the chip. The capacitance between the metal pad and the chip is about $\varepsilon A/d$, and so increased metal pad area can increase the capacitance. Right, the modulation response as a function of the device capacitance

$$\mathrm{BW}_{\mathrm{damping}}(\mathrm{GHz}) = \frac{9}{K(\mathrm{ns})}. \tag{8.36}$$

When the *K-factor* is extracted from a set of modulation measurements, it gives an estimate of what the maximum bandwidth for that laser can be. In the example discussed, where the *K-factor* turned out to be 0.25 ns, the maximum *K-factor* limited bandwidth is 36 GHz. At currents above that value corresponding to that bandwidth, the response is so damped that the total bandwidth is lower.

The second limit which is contained in the laser modulation equation is the 'parasitic' limit, which relates to the $1/(1 + j\omega\tau_c)$ term in the modulation equation. This equation represents a single-pole falloff and as such, the bandwidth associated with it is

$$\mathrm{BW}_{\mathrm{parasitic}} = \frac{1}{2\pi\tau_c}. \tag{8.37}$$

Hence, for the 10 ps capture time seen in the example, the bandwidth associated with it is about 15 GHz. This term is the easiest to engineer (either increase or decrease) and can be used to improve the laser response.

Those are the two fundamental limits but in practice, the device bandwidth can be limited by other empirical limits. The first of these to be discussed is the thermal limit. The bandwidth increases with increasing current, but increasing current also tends to increase the temperature of the device. At some point, this thermal effect put an end to the increases with current, and the modulation response saturates or even degrades when additional current is injected. The approximate maximum

Table 8.2 Limits to laser bandwidth

Limit (GHz)	Expression
K-factor limit	$\sim 9/K$ (ns)
Parasitic/transport limit	$1/2\pi\tau_c$
Heating limit	$\sim 1.5 f_{r\text{-}max}$
Facet power limit	Varies—typically 1 MW/cm^2 for uncoated devices

bandwidth due to this thermal limit is $1.5f_{r\text{-}max}$, where $f_{r\text{-}max}$ is the maximum observed resonance frequency.

There is a second limit sometimes imposed by the power-handling capacity of the facet. Higher bandwidths always require higher photon density, which implies a higher power density passing through the laser facet. The laser facet is a peculiarly vulnerable part of the laser. The atomic bonds on the facet are unterminated, and there are often defect states associated with them. These states can potentially absorb light, creating heat. If photons are absorbed going through the facet, portions of the facet can actually melt. The melted facet absorbs even more light, which leads to even more degradation. This can lead to catastrophic facet damage as shown in Fig. 10.20.

This catastrophic optical damage (COD) limit is typically around 1 MW/cm^2 for an uncoated facet. Coating the facet for passivation of the unterminated bonds, or to adjust the location of the magnitude of the peak optical field, can substantially increase the amount of power the facets can tolerate. Unlike the other limits, if approached, it typically terminates the useful life of a particular device and so should be taken as a specification for a maximum allowable optical power out or operating current.

Table 8.2 lists the expressions for the modulation frequency limit and the laser bandwidth.

With all of these different limits to small signal modulation, what is the limit for a given laser at a given temperature? The limit, of course, is the lowest of these, which varies from device to device. Typical bandwidths for conventional eight quantum well 1.3 μm devices designed for directly modulated communication are often well over 20 GHz at room temperature. These devices are fast. Nowadays, they are being put together in products that can modulate at 100 Gb/s through a combination of different modulation schemes and multiple lasers and wavelengths.

8.6 Relative Intensity Noise Measurements

We have shown how information about the physics inside the laser can be extracted from optical modulation measurements. It is a very powerful technique, but it does have some disadvantages. Primarily, the laser itself must be packaged in a way that allows for high speed testing. Typically, either the laser is fabricated in a coplanar configuration such that it can be directly contacted with such probe or it is mounted

on a suitable high speed submount. The modulation speed for plain laser bars, probed with a single needle as pictured in Fig. 5.8, is limited by the inductance of the needles to well under a GHz, and so the fundamental laser modulation speed cannot even be measured.

In addition, measurement of electrical-to-optical modulation includes terms like transport to the active region and capacitance that can obscure active region dynamics.

However, information about the high speed properties can be obtained through a simple DC measurement, from the laser relative intensity noise (RIN spectrum). The basic process and measurement technique are illustrated in Fig. 8.11.

The basic process is illustrated in the top sketch. A laser, above threshold, has the majority of its emission from stimulated emission. However, there is still background of random radiative recombination from spontaneous emission. This spontaneous emission at random times acts as a broadband noise source input into the laser cavity. This noise (primarily created by random recombination coupled into the lasing mode) is amplified by the laser cavity frequency response curve. The result is an equation for relative intensity noise,

$$|\mathrm{RIN}(f)| \prec \frac{Af^2 + B}{\left(f^2 - f_\mathrm{r}^2\right)^2 + \frac{\gamma^2 f^2}{(2\pi)^2}} \tag{8.38}$$

where the denominator looks very like the modulation expression. In fact, from a spectrum of relative intensity noise data, the dependence of resonance frequency on input current (the D-factor) can be easily determined and the damping factor γ can be sometimes extracted. The peak (seen in the RIN curve) is the same as the peak shown in the modulation response curve.

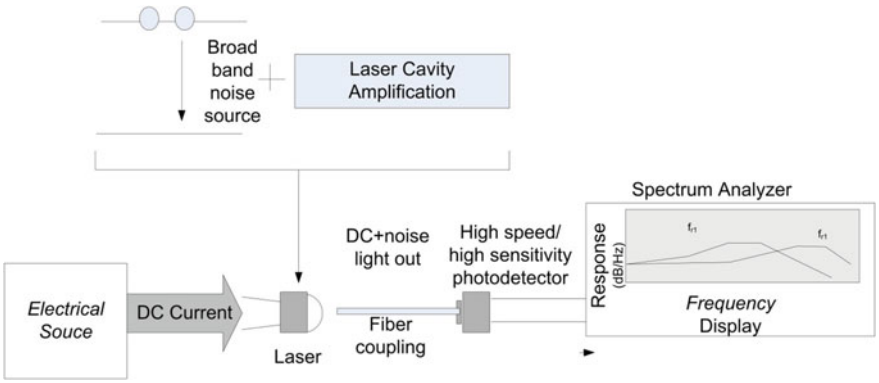


Fig. 8.11 Process and measurement of relative intensity noise. Random radiative recombination acts as a broadband noise source into the cavity, which then amplifies the noise in a manner similar to direct electrical modulation

This is a useful measurement technique even where directly modulated measurements are available, since it measures the characteristics of the cavity without external parasitics or the possibility of transport, or capacitance, influencing the dynamics of the device.

One pitfall is that it is a very sensitive measurement. Reflection between the fiber and the detector can show up as oscillations (spaced in the MHz) in the frequency signal if the fiber is not properly antireflection coated and the measurement done with insufficient optical isolation.

Relative intensity noise is a parameter that is sometimes specified in lasers, with requirements that it be less than values like −140 dB/Hz average, from 0.1 to 10 GHz,[2] at given operating conditions. Like electrical modulation, the RIN measurement peak increases with current and increases with device differential gain. Engineering the device for a high differential gain will move the resonance peak further to the right at a given current.

8.7 Large Signal Modulation

While the small signal bandwidth is of theoretical interest and includes much of the physics of the laser response, what is really relevant for most applications is the large signal response. For most digital modulation schemes, the relevant metric is the eye pattern which we introduced in the beginning of the chapter.

In an eye pattern measurement, binary data encoded as two (or multiple) current levels is driven into the laser, one representing a '0' (for example, 20 mA) and one representing a '1' (for example, 50 mA). These 1's and 0's occur in random patterns. The light out of the laser is measured with traces of all of them displayed. What is desired is a clear area with no signals in it, clean and sharp up and down transitions, and minimal overshoot and undershoot.

It is not obvious from laser characteristics, such as differential gain, what the eye pattern at a particular modulation speed will be, and yet it is important to tie the laser physics to the device modulation performance. This can be done using the versatile tools of the rate equations, which can be numerically solved to obtain the response for any input current.

8.7.1 Modeling the Eye Pattern

The rate equations do an excellent job of modeling the salient features of the small signal modulation response and can also be used to model the large signal response. In this case, the appropriate rate equations are the full rate equations in Eq. 8.15, not the small signal version. (Laser digital modulation is not a small signal!) The two rate equations for photon density and carrier density form a set of coupled nonlinear

[2]For example, this specification is from a Finisar S7500 tunable laser.

differential equations that can be numerically solved by a number of techniques, including the Runge–Kutta method (see Problem 8.4).

What this does is relates the small signal parameters to the large signal pattern (which is really of more direct interest). Figure 8.12 shows an example of a measured eye pattern and a simulated eye pattern obtained from numerical simulation of the rate equations using the parameters extracted from the small signal model.

As can be seen, it does a good job of reproducing most of the relevant features. The overshoot and the traces are clearly seen. With tools like this, the effect of changes in the *K-factor* or capacitance can be easily seen in the eye pattern. Optimization of the laser transmission can be more easily quantified.

The hexagon in the center and the shaded region on top of the measured eye pattern represent the eye mask, where traces from 1's and 0's are forbidden to cross. Typically, the quality of an eye pattern is determined by how far away the eye traces are from this forbidden region, measured in a percentage of 'mask margin'

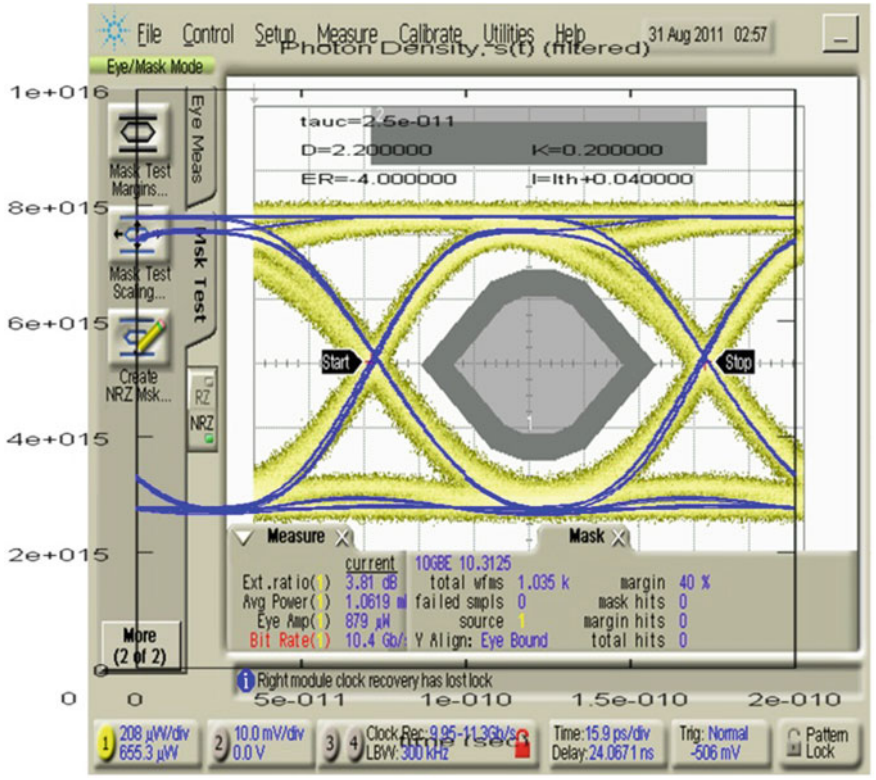


Fig. 8.12 Comparison of measured eye pattern with simulated eye pattern (thin lines). The parameters used in the simulation (dg/dn, ε, and the capacitance time constant, τ_c) are extracted from small signal analysis. The hexagon in the center and the shaded region on top represent the eye mask, where traces from ones and zeros are forbidden to cross. Typically, the quality of an eye pattern is determined by how far away the eye traces are from the forbidden regions (gray)

for a given device. There are different eye masks for different applications (including SONET and Gigabit Ethernet), and the required transmission characteristics also differ from application to application. During the measurement, the device is filtered by a low-pass filter with a bandwidth a little below relevant gigabit speed to suppress the inherent ringing and overshoot associated with all semiconductor lasers. For example, a 10 Gb/s receiver will often use an 8 GHz low-pass filter in front of the optical input data.

8.7.2 Considerations for Laser Systems

Before we leave the topic of laser transmitters, it is worth addressing some laser-in-a-package issues that are important to achieving a working transmitter system. A typical laser in a package is illustrated in Fig. 8.13. The package is a TO-can with a lens on the top. The cutaway view shows (not to scale) the laser mounted on a simple submount with metal traces. Also on the submount is what is called a back-monitor photodiode, which detects the light coming out of the back facet of the laser. Because the light out of the device varies enormously with temperature and slightly with aging, this allows the control system to adjust the current to the laser to maintain a more constant power into the fiber.

The driver, which is shown as a triangle in the diagram, is a high-performance piece of electronics that modulates high current sources at very high speeds. These speeds of 10 Gb/s or even more are well into the microwave regime of circuit design. Hence, the traces have to be designed for high speed signals and impedance-matched to the impedance of the driver. Wire bonds used to connect the driver to the TO-can, and the submount to the laser, have to be short.

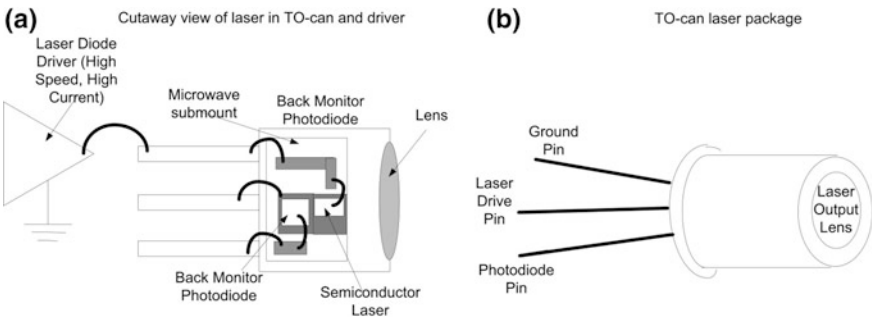


Fig. 8.13 **a** A cross-sectional view of a packaged laser system and laser, and **b** a sketch of the final packaged product

Optical issues are also important. Reflection back into the laser can lead to kinks in the *L-I* curve, mode hops, and deleterious behavior. Sometimes, laser packages are designed with optical isolators which prevent back reflection from reaching the laser, but low-cost transmitters often omit them.

8.8 Summary and Conclusions

In this chapter, the basics of direct modulation in lasers were covered. The use of eye patterns as metrics for directly modulated, digital transmitters are illustrated. Typical eye patterns from modulated lasers show inherent frequency effects due to the physics of the laser.

To understand these effects, we first analyze the small signal response of a laser. The rate equations are linearized, and the results show a characteristic oscillation frequency and decay time related to the photon lifetime, carrier lifetime, and operating point of the laser. This homogenous response has strong effects on the modulation response (with a sinusoidally modulated small signal current). The small signal frequency response is given and also includes the effect of the characteristic oscillation (resonance) frequency.

From small signal response measurements, fundamental characteristics of the laser active region can be extracted. These include differential gain, gain compression, and the equivalent parasitic capacitance associated with the device. These parameters, and particularly, the parasitic capacitance, can be engineered to improve the device performance for directly modulated communication.

The rate equation model, along with practical considerations, gives some limits to the small signal laser bandwidth. Both laser fundamentals (K-factor and parasitics) and operating issues (facet power handling and temperature issues) limit the bandwidth, and in general, the bandwidth is limited by the most restrictive of these.

These parameters can also be used to model the large signal response through numerical solution of the rate equations using laser parameters extracted from small signal measurements. This model can show how the operating point (high and low current levels) or parasitics affect the eye pattern of the device.

At the end of the chapter, a brief discussion of laser specifications, and of packaging, connects laser fundamentals to laser applications as communication devices.

8.9 Learning Points

A. Many communications lasers are designed for directly modulated digital transmission, and a clean difference between a low and high level is desired. However, overshoot and undershoot are inherently part of the laser dynamics.

B. Small signal modulation and the measured laser bandwidth are excellent and easily characterized metrics for large signal performance.
C. Small signal measurements can provide information about the fundamental physics of the laser active region.
D. Bandwidth measurements are made with a small signal superimposed on a DC bias, and the optical response at fixed input amplitude plotted versus frequency.
E. The frequency response of an LED is limited by the carrier lifetime.
F. The homogenous small signal response of a laser is a decaying oscillation, with both the oscillation frequency and the decay envelop both dependent on the bias point. The decay time of the homogenous small signal solution also depends on the carrier lifetime; the resonance frequency of the homogenous solution also depends on the geometric average of the carrier lifetime and photon lifetime.
G. To overcome these resonance frequency oscillations, typically the receiver is low-pass-filtered.
H. The modulation response function of a laser is the small signal variation of light out as the current is modulated (superimposed on a DC current) as a function of frequency.
I. The modulation response frequency of the laser is a second order function characterized by a resonance frequency and a damping factor, as well as a first order parasitic/capacitive term.
J. Typical analysis takes a set of modulation measurements at different bias conditions, from which the differential gain and gain compression factor can be extracted. A similar analysis can be done on relative intensity noise spectra with a DC drive current.
K. From the modulation equation, two fundamental limits to laser modulation frequency can be derived: a K-factor limit, based on how fast the resonance peak damps out as it moves out in frequency, and a transport/capacitance limit, based on the limit based on transport to the active region, and the RC laser diode characteristics.
L. The laser bandwidth may also be limited by power handling capacity of the facet or the thermal effects when high current is injected.
M. The parameters extracted from a small signal analysis, such as differential gain, gain compression, and K-factor, may be used to accurately model large signal modulation shapes.
N. Directly modulated laser packages are typically specified for wavelength, speed, extinction ratio, and launch power. From the specifications, the operating point can be determined.
O. The current high speed of direct modulated laser transmission means that package and driver electronics must also be designed to handle those frequencies (typically up to 25Gb/s in 2019).

8.10 Questions

Q8.1. What factors limit the bandwidth of an LED?
Q8.2. What limits the small signal bandwidth of laser? Would you expect a VCSEL with a cavity length of ~1 μm and a facet, reflectivity, 0.99 to have a better bandwidth than an edge-emitting device with a cavity length of 300 μm and typical reflectivity of 0.3?
Q8.3. What limits the bandwidth of a transistor? How are transistors fundamentally different from lasers in this respect?
Q8.4. In the diagrams of Fig. 8.5, the current is actually switched at $t = 0$ ps, but the light starts to switch at about 40–50 ps afterward. What is responsible for that delay?
Q8.5. What is the order of magnitude for maximum directly modulated laser frequency? Suggest some design considerations for a high speed device.

8.11 Problems

P8.1. Suppose the radiative lifetime for an LED is 1 ns and the non-radiative lifetime is 10 ns. Find the bandwidth of the LED and the radiative efficiency of the LED.
P8.2. Some of the expressions for carrier density include a photon density S. An uncoated semiconductor laser has the following characteristics: $\alpha = 40$/cm, $n_{\text{modal}} = 3.3$.

(a) Calculate the photon lifetime.
(b) The measured resonance frequency is 3 GHz. Calculate the differential gain when the laser has photon density of 2×10^{16}/cm^3 (Neglect the ε/τ term).

P8.3. A particular cleaved laser has the following characteristics:
$\lambda = 0.98$ μm, $dg/dn = 5 \times 10^{-16}$ cm^2, $\tau_p = 2$ ps, $n_{\text{modal}} = 3.5$.
It can tolerate a facet power density of 10^6 W/cm^2 before degradation, and its facet dimensions are 1 μm by 1 μm.

(a) What is the maximum facet power the device can put out before catastrophic facet degradation sets in?
Assume the internal photon density in the cavity is 1.2×10^{15}/cm^3 at this maximum power.
(b) What is the resonance frequency f_r of the cavity at this power level. Assuming the bandwidth = $1.5f_r$, what is the maximum bandwidth due to facet power capabilities?
(c) If the devices' *K-factor* is 0.9 ns, will fundamental or facet power limits determine the bandwidth?

P8.4. The objective of this problem is to numerically calculate the response of a laser which has been switched from one current value to another above threshold. This is very similar to how the laser would be used in a directly modulated setup.
The device in question has an active region volume of 120 μm^3, a photon lifetime $\tau_p = 4$ ps, $\tau = 1$ ns, $\beta = 10^{-5}$, $dg/dn = 5 \times 10^{-15}$ cm^2, $\varepsilon = 10^{-17}$ cm^{-3}, and $n = 3.4$.

a. Calculate the threshold current in mA.
b. Find the steady state value of n and s at $I = 1.1I_{th}$.
c. Using an appropriate technique, numerically calculate the response of the laser if the current is suddenly switched to $4I_{th}$ for 100 ps and then switched back to $1.1I_{th}$. This should look similar to the eye pattern response.

P8.5. We would like to expand the rate equation model we have, which is written in terms of carriers in the active and photon density, to also include carrier transport from the injected contacts and edge of the cladding to the active region. Shown in Fig. P8.14 is the diagram of the core, cladding, and active region. Write a third rate equation which features current being injected into the cladding, rather than directly into the active region, and includes the carrier transport time τ_c from the cladding to the core. Assume there is no transport from the core back to the cladding.

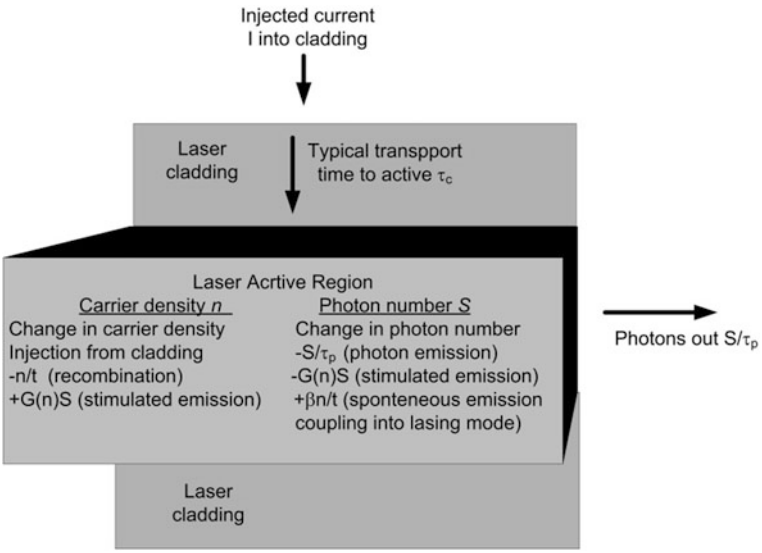


Fig. P8.14 A rate equation picture of a laser, including transport from the cladding to the active region

P8.6. Figure 8.10 shows the geometry of the extra capacitance induced between the contact metal pad and the n-doped surface of the laser wafer. If the metal pad is 300 μm long and 200 μm wide, calculate the oxide thickness to give a capacitance associated with the pad of 2 pF.

Check for updates

Distributed Feedback Lasers

9

...and there, ahead, all he could see, as wide as all the world, great, high, and unbelievably white in the sun, was the square top of Kilimanjaro.
—Ernest Hemingway, *The Snows of Kilimanjaro*

Abstract

Good-quality long-distance optical transmission over fiber needs lasers which emit at a single wavelength. This is almost universally realized by putting a wavelength-dependent reflector into the laser cavity, in a distributed feedback laser. In this chapter, the physics, properties, fabrication, and yields of distributed feedback lasers are described.

9.1 A Single-Wavelength Laser

The mountain top of Kilimanjaro, like the cleaved facets of a Fabry-Perot laser, reflects all colors. Though it may be 'great, high, and unbelievably white,' this *wavelength-independent* reflection means that wavelength emitted by the cavity is determined only by the gain bandwidth of the cavity and the free spectral range of the cavity. Because the reflectivity is wavelength independent, typically the emission from an edge-emitting Fabry-Perot device has many peaks in a range of 15 nm or so (see Fig. 9.1b).

What is needed for long-distance transmission, as we will talk about below, is a semiconductor laser whose optical emission spectrum is as narrow as possible. In this chapter, we describe how a semiconductor gain region gain can be made to emit in a single wavelength. The technology of choice for this (and the primary focus of this chapter) is the distributed feedback laser, usually abbreviated DFB.

© Springer Nature Switzerland AG 2020

D. J. Klotzkin, *Introduction to Semiconductor Lasers for Optical Communications*,
https://doi.org/10.1007/978-3-030-24501-6_9

9.2 Need for Single-Wavelength Lasers

By 'single wavelength,' what we mean is a device whose spectrum measured on an optical spectrum analyzer has one dominant wavelength, whose peak is typically 40 dB (10^4) higher than all the other peaks. This is illustrated in Fig. 9.1. Shown next to it, in comparison, is the output of a Fabry-Perot laser, which is composed of many peaks separated by the free spectral range and set by the gain bandwidth of the device. Other features of the spectra are labeled and will be discussed later in the chapter.

Single-wavelength lasers are important for three reasons. First, a principal use for communication lasers is direct modulation on fiber. In optical fiber, light at different wavelengths travels at slightly different speeds. This is called dispersion. The effect of dispersion on transmission is as follows: Suppose a current pulse is injected into a Fabry-Perot laser, causing the optical output power to change from one level (say, 0.5 mW) to another level (say, 5 mW). A detector in front of the laser will register a clean 'zero-to-one' transition. However, because this optical power will be carried by many different wavelengths traveling at different speeds, after a few tens or hundreds of kms down the fiber, the clean transition will be degraded. Eventually, a set of ones and zeros will be smeared out into a uniform level. The idea of pulse degradation as it travels because of dispersion is illustrated in Fig. 9.2. The pulse in the Fabry-Perot laser is carried by three wavelengths (for the sake of illustration); after kilometers of travel, the three wavelengths traveling at different speeds arrive at different times, and it is difficult to reconstruct the original data.

A good analogy of dispersion is the runners in a 26.2-mile marathon. With a wide enough starting line, all the runners can start at the same time, but they all run at different speeds. If they are only running a block, their finishing times will only be slightly different. However, if they are all going 26.2 miles, the faster ones will finish hours after the slower ones, and the sharp beginning of the race will have a lingering finish that is hours long.

If all the runners were picked to be about the same speed (analogous to having the light pulse all carried at one wavelength), the finish would be nearly as sharp as the start. A series of 'marathons' launched a few minutes apart would be distinguishable at the end of the race. The dispersion of the race is effectively low because the speed of the runners is nearly the same. In a single-wavelength laser, a

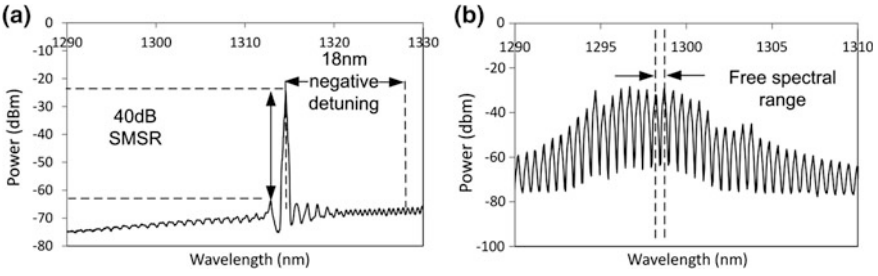


Fig. 9.1 Optical output spectra from **a** a single-mode, distributed feedback laser and **b** a Fabry-Perot, with some labeled features discussed in the text

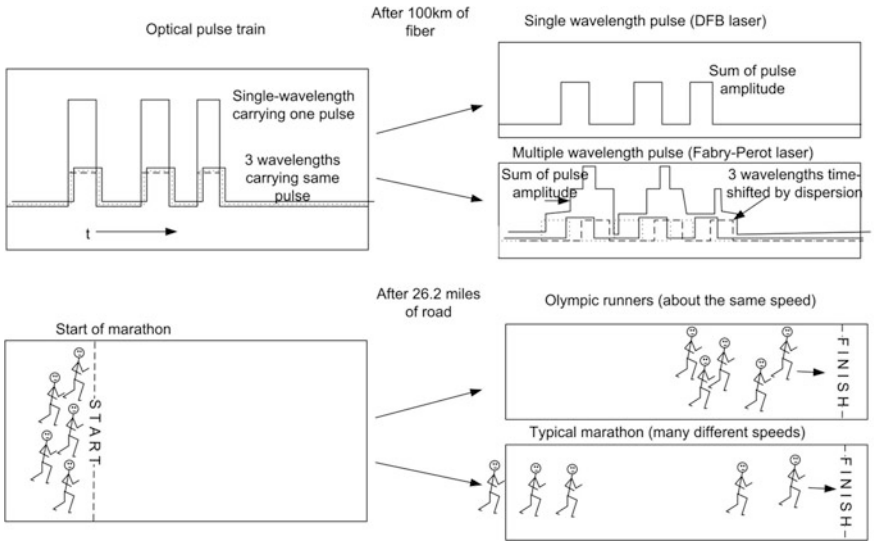


Fig. 9.2 Top, dispersion in an optical pulse train due to different speeds of light down a fiber; bottom, dispersion in finish times in a marathon due to different speeds of various runners. In order to clearly see ones and zeros after traveling many kilometers in optical fiber, the original source should be a single-wavelength device

pulse, once launched, can be resolved many kms later. This somewhat strained analogy is pictured in Fig. 9.2 and then (rightly) abandoned.

Though optical absorption is very significant over 100 km or more, it is less of a fundamental barrier because fiber amplifiers (like the erbium-doped fiber amplifier) can regenerate optical signals easily with near-perfect fidelity, though with the addition of noise. The fundamental limit to transmission is always signal-to-noise ratio. Dispersion is most important in the 1550-nm wavelength range where fiber loss is minimal. Around 1310 nm, dispersion is close to zero, but the loss is much higher. The 1550-nm wavelength range is what is used for long-distance transmission.

A second reason that single-wavelength lasers are important is bandwidth. Each fiber can transmit with reasonably low losses over at least 100 nm of optical bandwidth (from 1500 to 1600 nm); each 'channel' of modulated information is carried on a wavelength band in the fiber. This typical scheme is called 'dense wavelength division multiplexing' (DWDM). The narrower the channel, the more channels can be carried on a fiber. If each channel is <1 nm (typical of single-mode lasers), then more than 100 channels can fit on a fiber; if the channels are carried by Fabry-Perot lasers with optical linewidths >1 nm, the capacity of the fiber is much less. Additionally, distributed feedback laser wavelength is much less sensitive to temperature. This is important because in wavelength division multiplexing the wavelength has to be very well controlled in order to not interfere with channels on adjacent wavelengths.

Table 9.1 Necessity for a single-wavelength device

Property	Requirement
Dispersion	Light with different wavelengths travels at different speeds in a fiber. If the device is close to single wavelength, it can be more easily received after traveling many kms
Channel capacity	If each device is restricted to a narrow range of wavelengths, more devices can be carried on the same fiber
Speed/design degrees of freedom	Distributed feedback lasers can put the lasing wavelength away from the gain peak, leading to higher-speed devices and another degree of design freedom

Finally, there is one design feature of distributed feedback lasers which gives another degree of freedom in laser design and makes distributed feedback devices faster than Fabry-Perot devices. As will be seen, the lasing wavelength is set by the grating period and is independent of the gain peak of the material. If the lasing wavelength is shorter wavelength (higher energy) than the gain peak, the device is said to be negatively detuned. This negative tuning results in higher differential gain and a higher-speed device.

The benefits and need for these single-wavelength devices are summarized in Table 9.1.

Below, we discuss some other ways to achieve single-mode emission before exploring the distributed feedback structure.

9.2.1 Realization of Single-Wavelength Devices

Single-mode devices can be realized in a few ways, and before we discuss in detail distributed feedback devices, let us introduce some of the other methods that can be used.

9.2.2 Narrow Gain Medium

The simplest possible way to get a single wavelength is to have a gain medium that is very narrow, so that there is only optical gain in a small range. For example, HeNe and other lasers based on atomic transitions lase with very narrow spectral width and at a single precise wavelength. If there was only optical gain over a spectral range <1 nm, then clearly there would be an optical linewidth of <1 nm. Theoretically, that is certainly true, but practically speaking, gain regions composed of semiconductors cannot be made narrower than many tens of nm.

Even active regions based on quantum dots are several tens of nm wide, due to the size variation of the dots. Nonetheless, the overwhelming advantages of semiconductor lasers (small size, low power, high speed, and the ability to realized

useful wavelengths in the near-infrared range) outweigh the difficulty we will see in getting lasers to lase at just one wavelength.

9.2.3 High Free Spectral Range and Moderate Gain Bandwidth

From Chap. 7, we saw that putting the gain region into a Fabry-Perot cavity imposes a free spectral range (FSR) on the output of the device, pictured again in Fig. 9.1. This free spectral range increases as the cavity width decreases. Typical edge-emitting cavities of 300 μm or so have free spectral ranges of about a nm, and so, there are many peaks coming out of the cavity.

The formula for free spectral range $\Delta\lambda$ (adapted from Chap. 7) is

$$\Delta\lambda \approx \frac{\lambda^2}{2Ln_g} \tag{9.1}$$

where λ is the lasing wavelength, L is the cavity length, and n_g is the group index. If the free spectral range is much shorter than the cavity gain bandwidth, many lateral modes are possible.

However, suppose the cavity length was engineered to be less than 2 μm so the peak-to-peak spacing was greater than 20 nm, the typical gain bandwidth. In that case, there would only be one peak in the gain bandwidth, and the device would be single mode. Such a device exists. It is commonly made as a vertical cavity surface-emitting laser (VCSEL) and is illustrated (in comparison with a standard edge-emitting laser) in Fig. 9.3.

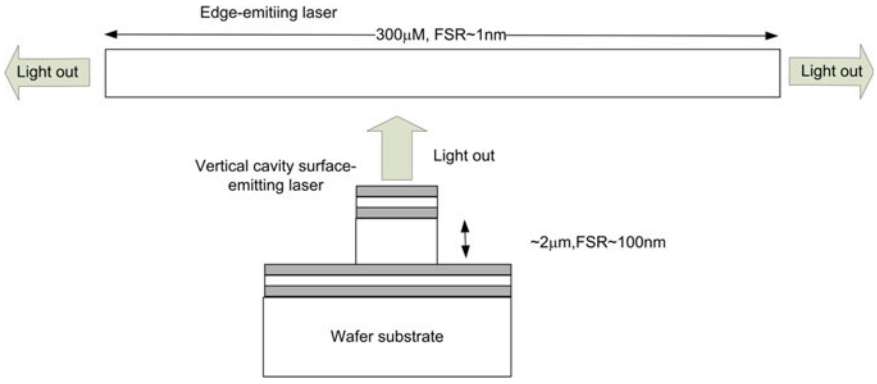


Fig. 9.3 Top, a sketch of an edge-emitting laser, with a 300-μm-long cavity and hence a very short free spectral range. This device can have multiple lateral modes and emits from the front (and back). Bottom, a VCSEL device which has a cavity length of a few microns, and hence a free spectral range of >100 nm, such that only one longitudinal mode is supported. The VCSEL emits from the top and bottom, and so its cavity length is about the thickness of the quantum well and cladding

Because the VCSEL cavity is so much shorter, the free spectral range is much larger. In fact, for a typical mirror-to-mirror VCSEL spacing of 3 μm, the free spectral range is >100 nm. The gain region, however, is the same as in a quantum well laser and about 10–20 nm wide. *Since the free spectral range is larger than the gain bandwidth, only one wavelength will fit within it, and these devices are inherently single (longitudinal) mode.*

However, VCSELs are not yet the solution for laser communications. The potential issues with these devices would easily make a chapter or book in themselves, but fundamentally they have two problems which make them unsuitable substitutes for edge-emitting lasers. First, because the gain region is very short, the mirror reflectivity is very high (to keep the optical losses low). This means that most of the photons created are kept within the VCSEL cavity, and the power output of a mW or so is not quite enough for fiber telecommunication needs. Second, the very short gain region means the device operates at a very high gain (and high current density) and so suffers from heating due to current injection. Typically, VCSELs do not operate over as high a temperature range as edge-emitting lasers.

There is another technological factor which makes VCSELs a better technology for shorter wavelengths than for the 1310- and 1550-nm wavelength devices. The very high reflectivity of VCSELS is realized with Bragg reflector stacks of materials of two different dielectric constants. It so happens that for GaAs-based devices (with wavelengths up to 850 nm or so) GaAs and AlGaAs form a very nice material system for these Bragg reflectors. In the InP-based system, it is not as easy to realize these Bragg reflectors on the top and bottom of the device.

Vertical cavity lasers do have a huge technical role in products like CD players and other low-cost, less demanding laser applications. They are lower cost than edge-emitting lasers and easy to test, but they do not have the necessary performance at the right wavelengths for high-quality fiber transmission.

9.2.4 External Bragg Reflectors

If we cannot reduce the gain bandwidth to below 10 nm and very short cavities are impractical, another alternative is to narrow the reflectivity range. Cleaved facets are largely wavelength independent, but if some sort of wavelength-dependent reflectivity could be coated in front of the cavity, that would introduce a wavelength-dependent loss, which might be sufficient to induce a single-wavelength emission.

This facet coating is done all the time commercially, just not for the purpose of wavelength selectivity. Commercial lasers do not generally get sold with ‘as-cleaved’ facets; typically, they are coated with a low-reflectance (LR) coating on one end and a high-reflectance (HR) coating on the other. The HR coating is typically a Bragg stack in which each material is $\frac{1}{4}\lambda$ thick, and consists of one, or a few, dielectric layer typically sputtered onto the facets of the laser bars. A typical recipe might be alternating layers of SiO_2 (n = 1.8) and Al_2O_3 (n = 2.2). The schematic realization of this is pictured in Fig. 9.4. These coatings change the slope

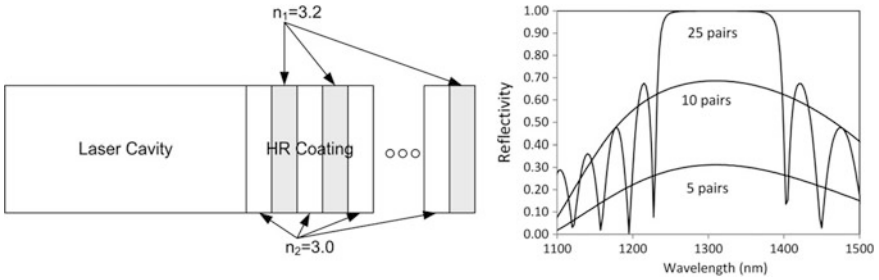


Fig. 9.4 A laser cavity with an external quarter-wave reflector stack and the calculated reflectivity as a function of the number of pairs. Potentially, the reflectivity can be higher than a cleaved facet, but typically, few periods of a high-contrast material are not very wavelength selective and have a broad reflectance band

asymmetry of the device and cause much more light to come out the end that couples to the fiber than the other end.

While this coating works very well for increasing the net reflectance, dielectric coatings composed of a few periods of materials with fairly high index contrast inherently have broadband reflectance across quite a range of wavelengths. Figure 9.4 shows a facet-coated laser and the calculated reflectivity as a function of the number of pairs of ¼-wavelength dielectric layers. (The reflectivity here as a function of wavelength is calculated using the transfer matrix method, which will be discussed in Sect. 9.5.)

Note that the reflectivity is fairly high over a wide region. While these dielectric stacks increase the reflectivity, they are no aid to wavelength selectivity.

Observing that this is what happens when a few periods of material with a relatively large index difference form the grating, we can calculate what happens when we have many, many periods of layers with a small dielectric contrast between them. The results of this are shown in Fig. 9.5. In this calculation, the refractive indices of the different dielectric layers differ by order of only 10^{-3}, and

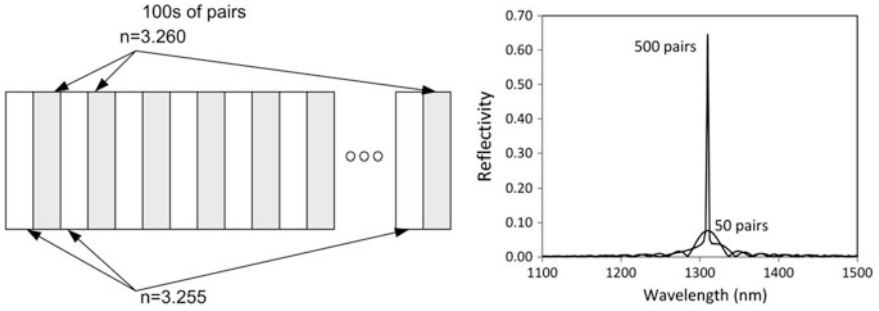


Fig. 9.5 Reflectivity of many pairs of dielectric layers with a low index contrast. The reflectance band is much higher, but the necessary thickness is 100 s of microns

so to get reasonable reflectivity from them, it is necessary to have many pairs. However, the reflectivity bandwidth is much, much narrower than that seen with fewer pairs of higher index contrast. Reducing the index contrast, n_1/n_2, with more pairs of dielectric levers dramatically narrows the reflectance band.

This is potentially promising, but there are important practical problems. A structure with 500 pairs of layers, each about 200 nm thick for maximum reflectance at 1310-nm wavelength, has about 100 μm of coating thickness. This is a very impractical thickness. For one thing, the light coming out of lasers is diverging and not collimated (see Fig. 7.11), and so that set of dielectric layers will not reflect 70% of the light back into the waveguide. It is also difficult to picture coating thicknesses of 100 s of μms on a 3-μm square facet. Mechanically, the coatings would be quite likely to peel off, crack, or otherwise fail.

9.3 Distributed Feedback Lasers: Overview

Finally, if a narrow gain bandwidth is impractical, a narrow cavity unsuitable for fiber transmission, and a Bragg reflector not useful, what is the solution? Figure 9.5 points the way to what has become the commercial single-mode laser method. If the number of periods is very high (a few hundred) and the index contrast is very low (less than 1%), the calculated reflectivity is very wavelength-specific, with a bandwidth of a few nanometers and a distinct peak. This suggests that a more effective method would be to integrate the reflector itself directly into the laser cavity.

In the following sections, we will start with a physical picture and qualitative overview of how a distributed feedback laser works, and then work into the important parameters in designing them (coupling constant κ, length L, reflectivity of the back facet R, and others).

9.3.1 Distributed Feedback Lasers: Physical Structure

Figure 9.6 illustrates what a multiquantum well, distributed feedback laser looks like. Somewhere, either above or below the active region, a grating is fabricated into the device. Because the optical mode sees an average index that extends out of the active region, it sees a slightly different index when it is near a grating tooth than when it is far away from a grating tooth. Hence, as the optical mode goes left or right in the cavity, it constantly encounters a change in index from when it is over a grating tooth, to when it is not over a grating tooth, to when it is over a grating tooth again.

The optical model of a grating built into a laser cavity is shown in Fig. 9.6. The key is that there is a very low index contrast between the toothed and the

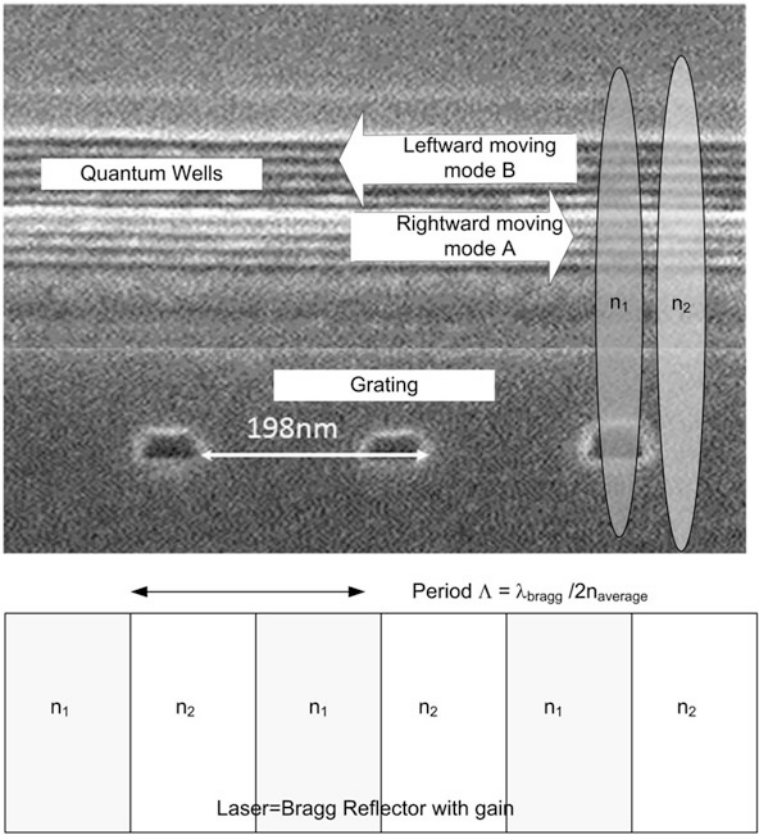


Fig. 9.6 Top, an SEM of a DFB laser showing the quantum wells and the underlying grating. Bottom, the optical model of the laser; the many, many periods of slightly different effective indexes serve as a wavelength-specific Bragg reflector

non-toothed regions. Typically, their effective index difference is about 0.1% or less. Because of that, the reflectivity model looks like Fig. 9.5 rather than Fig. 9.4.

As a prelude to the mathematical discussion that will follow in Sect. 9.6, the two counter-propagating modes 'A' and 'B' are also illustrated in the figure. Optical mode 'A' moves to the right; every time it encounters a grating tooth, a little bit of it is reflected in the other direction and joins mode 'B', moving to the left. Similarly, the left-moving mode 'B' is reflected just a bit at each interface and reflected in the 'A' direction, and modes 'A' and 'B' are said to be *coupled* together by the grating. This distributed reflectivity takes the place of mirrors on the facet and in addition introduces the exact right degree of wavelength dependence into the reflectivity.

9.3.2 Bragg Wavelength and Coupling

Two parameters used to characterize DFB lasers are the Bragg wavelength, λ_b, and the distributed coupling, κ. The Bragg wavelength, λ_b, defined in the figure above, is simply the 'center wavelength' of the grating defined by the grating pitch, Λ, and the average effective optical index n in the material.

$$\Lambda = \frac{\lambda_{bragg}}{2n} \tag{9.2}$$

At the Bragg wavelength, λ_{bragg}, each grating slice is $\lambda/4$ thick in the material. In a passive reflector cavity, the Bragg wavelength would be the wavelength of maximum reflectivity.

The coupling of a distributed feedback laser is characterized by the reflectivity per unity length. If n_1 and n_2 are the effective indexes that the modes see at those two locations, the reflectivity at each interface is

$$\Gamma = \frac{n_1 - n_2}{n_1 + n_2} = \frac{\Delta n}{2n} \tag{9.3}$$

where Δn is the slight difference between the modes of the effective indices and n is the average index. It experiences this reflection twice in each period Λ, and so the reflectivity/unit length is about

$$\kappa = \frac{\Delta n}{n\Lambda} \tag{9.4}$$

Because distributed feedback lasers are fabricated in various lengths, the usual parameter used to compare reflectivity is not κ, but the product κL (the product of reflectivity per length multiplied by the effective length). *This dimensionless quantity κL can be thought of as the equivalent of mirror reflectivity in a Fabry-Perot device*.

In general, the higher the κL is, the lower the threshold and slope become.

The Bragg wavelength λ_b is controlled by setting the period of the grating. Typically, a grating period of about 200 nm corresponds to a central wavelength of 1310 nm in most InP-based structures. The coupling κ is controlled by changing the strength of the grating, by moving it either closer or farther away from the optical mode, making it thicker or thinner, or change the composition to adjust the two effective indices, n_1 and n_2.

9.3.3 Unity Round Trip Gain

Just like Fabry-Perot lasers, there are two fundamental conditions for lasing in distributed feedback lasers:

(a) Unity effective round trip gain:
At the lasing condition, a round trip of the optical mode including lasing gain, loss through the facets, and absorption should lead back to the same amplitude as the original mode.

(b) Zero net phase:
Over the complete interaction with the cavity, the returning mode should be exactly in phase with the starting mode for coherent interference. It does no good to have maximum reflection at a particular wavelength that gets back to the starting point 180° out of phase.

In the next several sections, we will cover the math which describes distributed feedback lasers and shows how these conditions are met. Here, we present a more qualitative overview.

In a Fabry-Perot laser, changing the reflectivity of the facets changes the lasing gain of the cavity. The more reflective the facets are, the more the light is contained within the cavity, and the lower the threshold gain and threshold current. Introducing a grating into the cavity also changes the effective reflectivity with the advantage being that it does it in a very wavelength-dependent way.

However, *it is absolutely not as simple as the laser now lasing at the Bragg peak of maximum reflectivity*. The Bragg wavelength of maximum reflectivity is not necessarily the laser wavelength for minimum gain. This is counterintuitive, but true. If the light is created internally (as in a laser), the same interference effects that create reflection forbid the optical mode to propagate. There is a compromise between reflectivity and interference which moves the lasing gain minimum off the Bragg peak.

9.3.4 Gain Envelope

A more quantitative way to show this same point is shown in Fig. 9.7, which shows the calculated lasing gain envelope as a function of wavelength for the two different cavities of different κL, with typical laser absorption parameters. (This same graph for a Fabry-Perot laser would be a wavelength-independent straight line. The calculation method here is the transfer matrix method, which will be discussed in Sect. 9.5). As shown, for a fairly low κL device (with $\kappa L = 0.5$) the position of minimum gain is located at the Bragg peak; for a higher κL device ($\kappa L = 1.6$), the positions of minimum gain are symmetrically located around the Bragg peak.

In general, $\kappa L \sim 1$ are typical of index-coupled distributed feedback lasers.

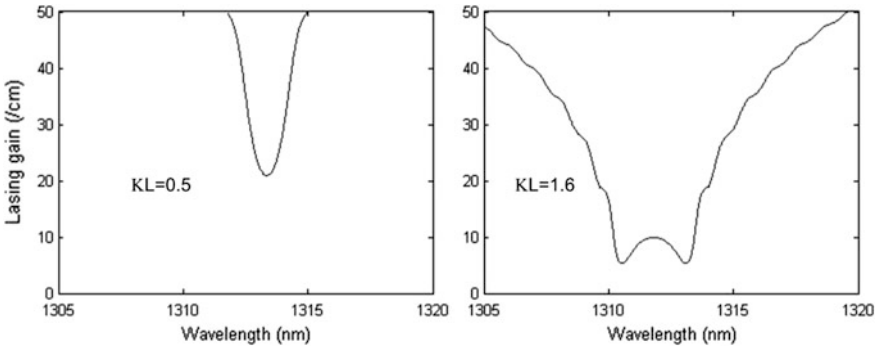


Fig. 9.7 Calculated gain curves for two different laser cavities, one with a low κL of 0.5 (left) and one with a high κL of 1.6 (right). The minimum gain is at the Bragg peak for the low κL cavity and at two symmetric locations outside of the Bragg peak for the high κL cavity

Although a higher κ (corresponding to a lower reflectivity) has a lower gain point, as κ gets higher, the minimum gain point drifts from the maximum reflectivity point. The critical difference between a distributed feedback laser and a Bragg reflector is that the Bragg reflector reflects external light that is incident upon it by creating destructive interference for light of a particular wavelength band inside the reflection surface. The light cannot propagate into the structure, and so it is reflected. In a distributed feedback laser, the reflector *is* the cavity. The light has to propagate somewhat to experience the necessary laser gain. The effect of the grating is to make the necessary lasing gain very dependent on wavelength.

9.3.5 Distributed Feedback Lasers: Design and Fabrication

The conditions for lasing for a DFB laser are exactly the same as in a Fabry-Perot laser: namely, unity round trip gain and zero net phase. Typical DFBs have one facet anti-reflection coated (as close to zero reflection as possible) and the other facet high-reflection coated, to channel most of the light out the AR-coated front facet. The zero net phase in a round trip is crucially affected by what is called the 'random facet phase' associated with the high-reflectivity back facet. That comes from the fabrication process for typical laser bars. In order to discuss this meaningfully, let us first briefly outline the fabrication process for a commercial distributed feedback laser.

We feel it is more productive to ease into the mathematics with a qualitative description first and so choose instead to dive directly into the conventional AR/HR DFB laser structure and its associated complications. In Sect. 9.6, we will discuss coupled mode theory which will give another way to look at these devices.

The typical process of turning a distributed feedback wafer into many bars of distributed feedback lasers is illustrated in Fig. 9.8. There are some important extra considerations above those required for a Fabry-Perot laser. The starting point is a

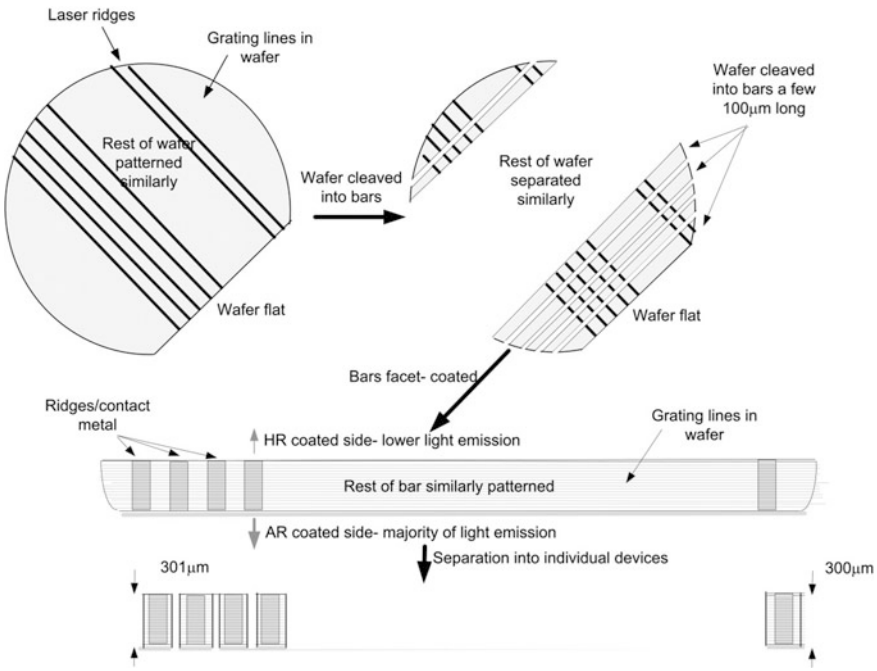


Fig. 9.8 Fabrication of DFB laser process, showing the origin of the random facet phase. The cavity thickness can vary slightly along the length of the bar, and variations on the order of a few tens of nm change the phase of the reflected light

wafer which has a grating already fabricated in it, along with all the rest of the necessary contact and compliant metals and dielectric layers. The wafer is then mechanically cleaved into bars which define the cavity length. Typical cavity lengths are usually hundreds of μms or so.

The gratings can be defined on the wafer in a holographic lithography patterning process, in which one exposure patterns lines of the necessary period on the whole wafer, or, alternatively, written by electron beam lithography. In recent years, electron beam lithography systems have developed the throughput to enable them to be used to write wafer-level grating patterns. However, without the ability to control cleaving with nanometer precision, a random facet phase is introduced.

After separation into bars, one facet is anti-reflection (AR) coated, and the other facet is high-reflection coated. The anti-reflection facet has reflectivity of <1%; it is designed to make the loss in the Fabry-Perot modes very high and ensure that the device only lases in the mode defined by the grating.

The anti-reflection coating in front is absolutely essential to get a good single-mode device. If it is missing, the lasing gain for the distributed feedback

peak and Fabry-Perot peak will be comparable and the laser could lase at a variety of wavelengths. Recall the lasing gain in a Fabry-Perot laser is

$$g_{\text{lasing}} = \alpha + \frac{1}{2L}\ln\left(\frac{1}{R_1 R_2}\right) \tag{9.5}$$

where L is the cavity length, α is the absorption loss, and R_1 and R_2 are the facet reflectivities (which are at most only weakly wavelength dependent). If R_1 or R_2 is very small (anti-reflective), the Fabry-Perot lasing gain g_{lasing} becomes very large, and the laser will lase at the mode defined by the grating. Fabry-Perot lasers are usually facet-coated also with the objective of increasing the power out of the front facet, but if that coating is missing, the result is simply a device with not as much power emitted out the front facet.

The number of grating lines can differ from device to device across a bar because it is impossible to pattern and cleave the device completely accurately. This causes a random facet phase associated with the high-reflectance facet that will be discussed next.

9.3.6 Distributed Feedback Lasers: Zero Net Phase

The wafer is cleaved into bars a few hundred microns long. The grating direction is in the same direction as the cleave direction (and perpendicular to the ridge direction) as shown in Fig. 9.8. The cleave, which is a mechanical operation, does not pick out an integral number of grating periods. Typically, there is a random residual fraction of a grating period left over. This does not matter on the anti-reflection side, because the light from that side is not reflected back into the laser cavity; however, it does matter very much on the high-reflection side.

A round trip through the Fabry-Perot cavity is required to have zero net phase, so that the round trip light undergoes constructive interference. The same is true in a distributed feedback laser; although the feedback is distributed, the net round trip length has to be an integral number of wavelengths. Distributed feedback lasers, like Fabry-Perot lasers, also have a comb of allowed modes set by the cavity length.

The random cleave at the end adds a certain random facet phase to the entire optical mode and shifts the set of allowed modes by a certain amount. Though the spacing may be the same, set by the length of the cavity, this random facet phase shifts all the points back and forth along the spectrum.

This random facet phase has great influence on the device operation. For a start, look at Fig. 9.9, which examines the net reflectivity from the highly reflective back facet with a small varying cleave distance remaining. The reflectivity of the back facet is the same: However, consider the reflectivity from the reference plane indicated on the diagram. In the first diagram, with no additional cleave length, the reflectivity is simply R. In the second, the reflected wave at the reference plane has an additional phase associated with the propagation of the left-going wave from the reference plane to the back facet and then back again. In the final case, the extra

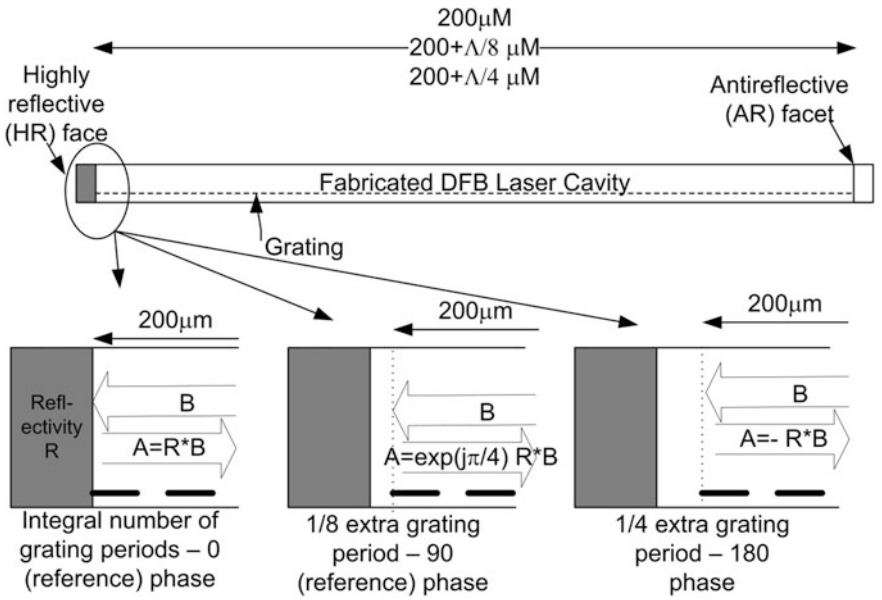


Fig. 9.9 A fabricated conventional DFB structure, showing the cause of the random facet phase and how it influences the effective reflectivity from the back facet

distance is sufficient to induce a 180° phase shift, and the reflectivity becomes $-R$. The magnitude of the reflected wave is always R, but the phase varies with the exact length of the laser in the typical HR-/AR-coated device.

Figure 9.10 shows with points the allowed lasing wavelength for a device with a particular length with two different back facet phases (indicated by dark and light points). The spacing between the allowed wavelengths is set by the length of the

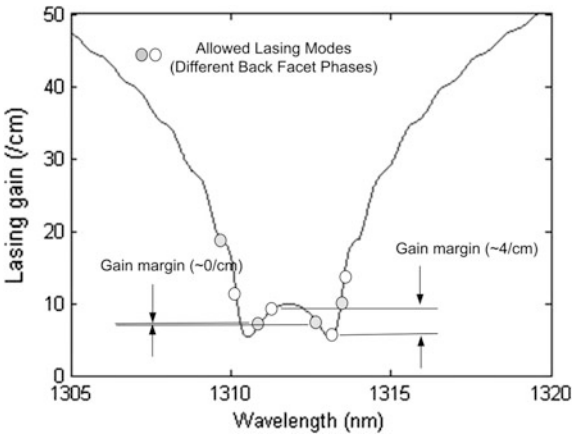


Fig. 9.10 Compared to a device with an arbitrary zero phase, whose allowed lasing modes are shown in white, a slightly longer device (whose allowed modes are darker) has its allowed modes by a fraction and may change the lasing mode dramatically

cavity just like a Fabry-Perot device and is about one nm for a cavity length of 200 μm. The random net phase comes from the random variation in cavity length from device to device.

In a Fabry-Perot device, this slight variation in cavity length does not do very much to the output. Slight variations in the length mean the device will shift its comb of allowed modes a bit, but the device will still lase at the allowed mode with maximum gain (which may shift by a fraction of a nm or so).

In a distributed feedback laser, these small shifts are extremely significant. When the allowed modes are shifted by a nm or two, the particular mode with the lowest gain can change dramatically. Figure 9.10 shows a device that would originally lase at the lowest gain point of ~1313 nm, shown by the lowest of the white dots. If the back facet phase were slightly different, it could lase near the other minimum at 1311 nm. Even worse, some other phase shift could leave two brown dots effectively at the same lasing gain (as illustrated). This would leave two allowed modes with essentially the same optical gain and lead to a device with two lasing modes.

Later on, we will talk about single-mode yield for distributed feedback lasers in the context of back facet phase, but qualitatively, the fundamental distributed feedback structure for index-coupled lasers usually has two symmetric points on the gain envelope, and the back facet phase determines where on the gain curve the device will lase. If two points are near the same gain, they may both lase, and it will not be a single-wavelength device.

Things get worse. Fabry-Perot lasers have a very simple power distribution inside the cavity, where the power is minimum in the middle and maximum at the ends. In distributed feedback devices, the power distribution also depends sensitively on the back facet phase, and so the slope efficiency out the front of the device varies with facet phase. Because the actual lasing gain also depends on the back facet phase, and the threshold current depends on the lasing gain, these as well vary significantly from device to device. These dependencies are listed qualitatively in Table 9.2.

Essentially, we have significantly improved over a Fabry-Perot, from a comb of modes spanning 10 nm or more to potentially one or at most two degenerate

Table 9.2 Effects of back facet phase on laser properties

Property	Explanation
Threshold current	Back facet phase affects the allowed lasing wavelengths which have different lasing gains
Lasing wavelength	Random back facet phase shifts the allowed modes slightly, but, since the gain varies significantly with slight wavelength changes, the mode with lowest gain can vary significantly (from one side to the other of the Bragg wavelength)
Single-mode behavior	With some back facet phases, two allowed modes have essentially the same lasing gain. In that case, the device can have two lasing modes
Slope efficiency	The power distribution in the device depends sensitively on the phase. Slightly different back facet phases mean different slope efficiencies. Unlike a Fabry-Perot, the slope efficiency depends sensitively on the back facet phase

distributed feedback modes. In practice, random facet phase and the gain curve of the active region often make the device lase in a single mode. The statistics of how the random facet phase affects device characteristics will be illustrated in Sect. 9.4 using a model and experimental data from a population of devices.

9.4 Experimental Data from Distributed Feedback Lasers

9.4.1 Influence of κ on Threshold Current and Slope Efficiency

The first thing to show with distributed feedback devices is the role of κ and κL in affecting the DC characteristics of threshold current and slope efficiency. The distributed reflectivity plays a role very similar to that the reflectivity R in Fabry-Perot devices. To remind the reader, as R increases, the optical loss decreases according to Eq. 5.3 which decreases the threshold current. The slope efficiency increases according to Eq. 5.17.

The same effect is seen with varying κL. This is beautifully illustrated in Fig. 9.11, which shows slope efficiency and threshold current for devices of the same size fabricated on the same wafer engineered by e-beam written gratings with varying κ. Data like this is gold!

As can be seen, as κ increases, both slope efficiency and threshold current decrease.

Similar to picking R to adjust threshold and slope, κ and κL strongly affect threshold and slope in distributed feedback devices. For high-power devices requiring high slope, κ is kept low. For high-speed devices that are typically short, κ is high to reduce optical loss.

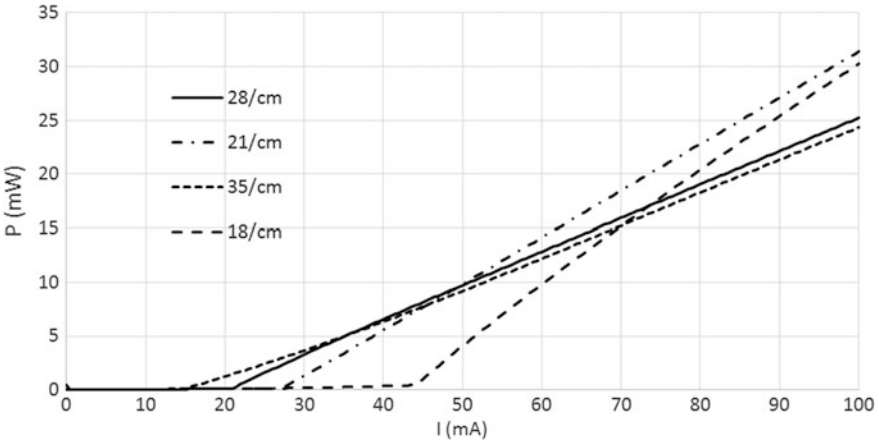


Fig. 9.11 *L-I* curve of a set of identical devices fabricated with varying κ, showing that the threshold current increases and the slope efficiency increases with reducing κ

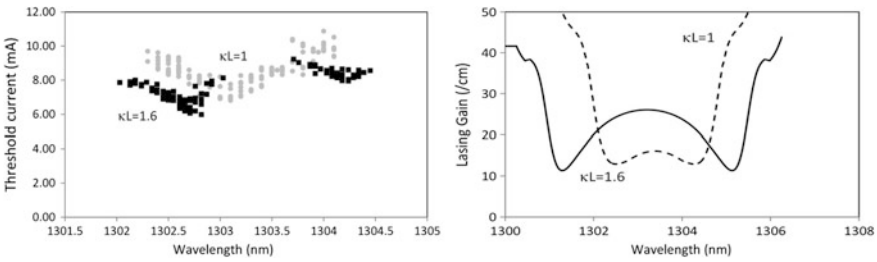


Fig. 9.12 Left, measured threshold currents of populations of nominally identical devices with random back facet phase, for two different populations with different grating strengths and kL values. Right, calculated lasing gain curves for the same kL. The shape of the measured threshold versus wavelength curve qualitatively matches the shape of the gain curve versus wavelength. The quantitative difference in threshold is not that high, because much of the threshold current is really transparency current

While average data varies substantially depending on κ, for devices of nominally the same κ, the random variation of back facet phase imposes a distribution on these nominal characteristics. This effect is the subject of the next sections.

9.4.2 Influence of Phase on Threshold Current

The fabrication of a distributed feedback device leads to a random phase on the back facet, and that random phase leads to variations in the laser properties.

The nice thing is that a single wafer, which typically has thousands of devices on it, has all the information needed to show these properties. When a population of lasers is fabricated, typically at some point they are cleaved to nominally the same length. However, the length of course cannot be controlled to the 100-nm scale with mechanical cleaving; therefore, the population effectively is of devices of nominally the same design, except with a random back facet phase.

Figure 9.12 shows the threshold current of two populations of identical devices (other than random back facet phase) with different κL, along with the calculated gain curve envelope. The lower the gain curve, the lower the threshold current is expected to be. No points are shown in the middle of the lasing band for the higher κL structure because there are no good single-mode devices in the middle of the lasing band for high κL structure. The variation in characteristics is due to the random back facet phase. In these graphs, the easily-measured lasing wavelength serves as a proxy for back-facet phase.

9.4.3 Influence of Phase on Cavity Power Distribution and Slope

The influence of phase on the output slope efficiency is not intuitively apparent. Starting with a calculated lasing gain and back facet reflectivity, the distribution of

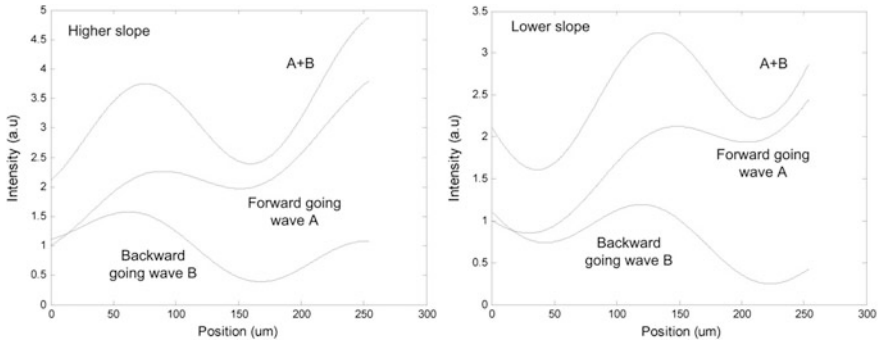


Fig. 9.13 Power distribution shown with two different back facet phases and hence different slope efficiencies, out of the cavity and power distribution within the cavity

power can be calculated throughout the laser cavity using the known gain. If the front facet is anti-reflection coated, as is usual, the relative slope efficiency will be proportional to the forward-going optical power intensity at the front facet.

Figure 9.13 illustrates this. Two different power distributions are shown, calculated for different back facet phases but otherwise identical laser structures.

There are several interesting things to be seen in these plots. First, notice that the total power density (forward plus backward) varies significantly inside the cavity and is not necessarily a maximum at the output facet. In contrast, Fabry-Perot devices always have the maximum optical power density at the facets. There is also a significant difference between the maximum and the minimum optical power distributions in these devices. This can cause subtle problems in device operation. Devices with strong difference between maximum and minimum power distributions are susceptible to *spatial hole burning*, where the carrier distribution is also not uniform because it is depleted by the large local photon density.

In the cavity on the left, with one back facet phase, the forward-going wave has an amplitude of 3.5, while for the one on the right, the forward-going wave has an amplitude of <2.5. The output slopes for these two devices will differ by more than 30%. Figure 9.14 shows measured and modeled slope efficiences for populations of devices with different κL showing ranges of observed slope efficiencies.

The pictures also show how the forward- and backward-going waves relate to each other. In a Fabry-Perot device (see Fig. 5.1), the forward-going wave grows as the backward-going wave shrinks, going toward one facet. Here, the backward- and forward-going waves grow and shrink together, because they are coupled to each other.

9.4.4 Influence of Phase on Single-Mode Yield

As seen in Fig. 9.10, the back facet phase particularly determines what wavelength the device lases at, by shifting the allowed modes on the gain curve envelope. Relatively, small shifts in back facet phase can change the mode with minimum gain significantly. Another consequence of the sensitivity of the lasing wavelength

to back facet phase is that it is quite possible to have two modes which have essentially the same lasing gain.

Figure 9.1 shows the usual metric for single-mode quality, the side-mode suppression ratio. The side-mode suppression ratio (SMSR) is the power difference between the highest power mode and the second highest power. Typically, the specification for a good single-mode laser is a side-mode suppression ratio of at least 30 dB.

The side-mode suppression ratio (SMSR) of a device depends on the gain margin for the device, where 'gain margin' means the difference between the lasing gain required for the mode with the lowest gain and the mode with the second lowest gain.

If the lowest mode has significantly lower gain required to lase than the second lowest mode, after the carrier population has reached the required lasing gain, it will be clamped; the carrier population will no longer increase with increase in current, and the device will lase only in that mode. If there are two modes which lase at about the same gain value on the DFB gain envelope, then it is possible that a given carrier density will be sufficient to support lasing in both modes. In that case, the output spectra of the device will have two prominent wavelengths. This is especially true due to the feedback mechanism of spectral hole burning, in which a high optical power density at one wavelength depletes carriers at that wavelength.

Hence, for a good single-mode device, it is required that there be sufficient gain margin between the two lowest lasing modes. Figure 9.15 illustrates a comparison of measured SMSRs along with calculated gain margin profiles for two different devices with different κL. The left side shows the measured side-mode suppression ratio of populations of devices, while the right side shows the calculated gain margin between the lowest and the next lowest modes.

Typically, gain margins of about 2/cm are needed for a good single-mode device. The gain margin for the two different phases is illustrated for the two different back facet phases in Fig. 9.10.

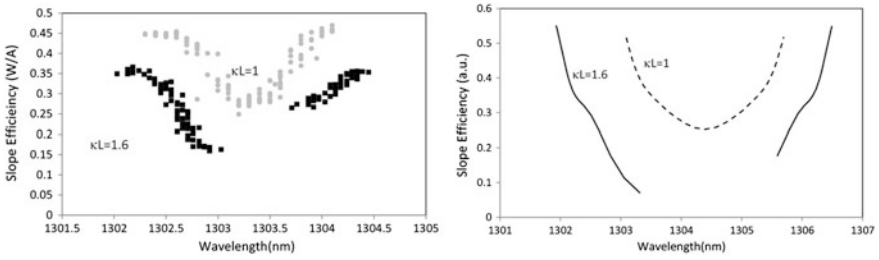


Fig. 9.14 Left, measured slope efficiencies of populations of nominally identical devices with random back facet phase, for two different populations with different grating strengths and κL values. The shape of the measured slope efficiency versus wavelength curve qualitatively matches the shape of the calculated slope efficiency curve versus wavelength. As can be seen, there is at least a factor of two differences in slopes from devices at the edge of the lasing band and those in the middle of the lasing band

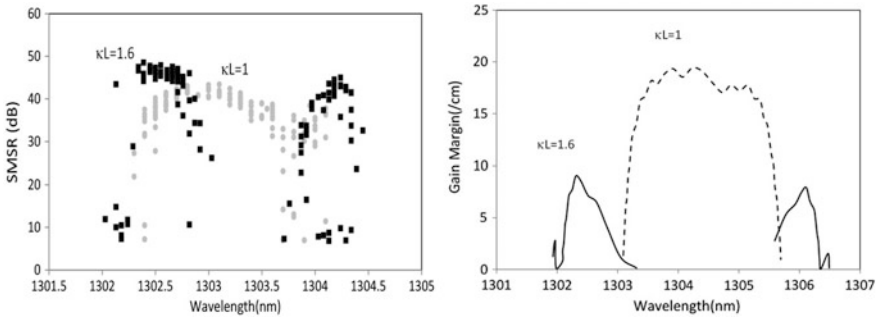


Fig. 9.15 Left, measured side-mode suppression ratio of populations of nominally identical devices with random back facet phase, for two different populations with different grating strengths and κL values; right, the calculated gain margin or difference between calculated gain of lowest mode and the next lowest mode. There is good qualitative agreement showing that for this device length, the higher κL material only had a good gain margin at the edges of the lasing band. In the center, the SMSR was low, and devices were multimode

For the high κL device, not only does the slope efficiency become minimal toward the Bragg wavelength, but the gain margin also becomes much lower. The devices close to the middle of the stopband tend not to be single mode, but multimode.

The point of these examples is to illustrate the significant influence of the random back facet phase on the lasing characteristics of otherwise identical lasers. Simply because the back facet phase varies randomly, some lasers will fail the specification typically due to low slope, poor SMSR, or poor threshold current. Values of κL determine not just the average static characteristics but the wafer yield.

The general effect of κ and κL on device properties is similar to what increasing reflectivity would be in a Fabry-Perot laser, decreased *threshold current* and *SE*. Effects on yield and such are more subtle.

Example: A typical laser has κL values about 1. Find the period and Δn, for a laser cavity 300 μm long with a κL about 1 designed to lase at about 1310 nm and an average mode index of 3.4.

Solution: If the target wavelength is 1310 nm, that means the Bragg wavelength of the grating should be targeted for 1310 nm. Hence, the grating period is

$$\Lambda = 1310\,\text{nm}/2/3.4 = 192.6\,\text{nm}.$$

For $\kappa L = 1$, κ (for a designed length of 300 μm) is 33 cm^{-1}, or

$$33 = \frac{\Delta n}{(3.4)192.6 \times 10^{-7}} = 0.0022,$$

or a change in index from one part to another of about 10^{-4}. This change in index is achieved by changing the structure (as shown in the micrograph in Fig. 9.6). The effective indices, n_1 or n_2, can be calculated through the methods in Chap. 7 or, more usually, calculated using finite-difference time-domain technique and numerical software.

Generally, the initial grating period and design are made based on calculation on models. Initial results are used to fine-tune the model and hit the precise wavelength in subsequent fabrication runs, as illustrated in the next example.

Example: The previous design is fabricated, but the average lasing wavelength turns out to be 1300 nm, not 1310 nm. Assuming the reason is that the calculated average effective index is off (but the laser layer structure stays the same) how would the design be altered in the next iteration to get 1310 nm?

Solution: If the actual wavelength turned out to be 1300 nm, then the effective index can be calculated from the same equation, as

$$192.6\,\text{nm} = 1300\,\text{nm}/2/n$$

which gives

$$n = 3.375.$$

Assuming n is 3.375, then the required grating period is

$$\Lambda = 1310\,\text{nm}/2/3.375 = 194.1\,\text{nm}.$$

In the second iteration, the target grating period should be 194.1 nm. Notice how precise the grating period has to be to get the wavelength to the target. Typical specifications for wavelength division multiplexed devices are within a nm; for wavelength tolerance like that, the grating period has to be specified, and accurate, to within 0.1 nm.

9.5 Modeling of Distributed Feedback Lasers

Let us spend a page or two to give a framework by which the statistics of different distributed feedback laser structures can be calculated. The specific details of the modeling are left as a problem at the end of the chapter.

The transfer matrix method for optical modeling is a general technique and is very good for modeling thin-film filters as well as distributed feedback lasers. The basic method is illustrated in Fig. 9.16, using the simplest optical example (propagation through a uniform dielectric). In the most general case, there is a left- and a right-propagating wave on both the left and the right sides of an arbitrary dielectric boundary with a refractive index, n_1, and a gain, g. The length of this dielectric we will set as $\Lambda/2$ (half the grating period) so that this small chunk represents one grating tooth.

The equations that relate the left and right sides to each other are

$$a_r = a_l \exp\left(g + j^{2\pi n_1}/\lambda\right)\Lambda/2 \tag{9.6}$$

$$b_r = b_l \exp\left(-g - j^{2\pi n_1}/\lambda\right)\Lambda/2 \tag{9.7}$$

We want to be able to write the waves on the right as a function of the waves on the left, so after some rearrangement, we can write

$$\begin{bmatrix} a_r \\ b_r \end{bmatrix} = \begin{bmatrix} \exp((g + j^{2\pi n_1}/\lambda)\Lambda/2) & 0 \\ 0 & \exp((-g - j^{2\pi n_1}/\lambda)\Lambda/2) \end{bmatrix} \begin{bmatrix} a_l \\ b_1 \end{bmatrix} = M_1 \begin{bmatrix} a_l \\ b_1 \end{bmatrix} \tag{9.8}$$

This expression has the 'output' (the waves on the right) as a function of the input (the waves on the left), times the transfer matrix M_1.

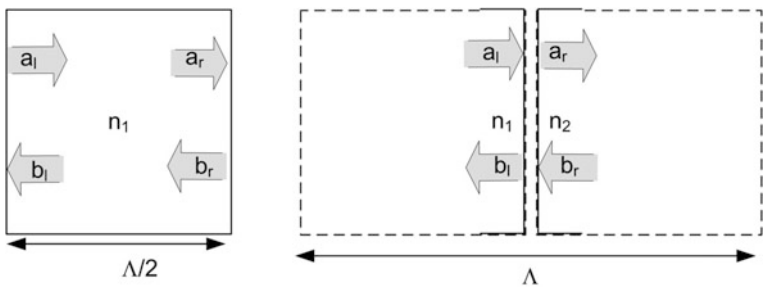


Fig. 9.16 Illustration of the transfer matrix method for light propagating in a region of index n_1 and n_2

In the second scenario pictured in Fig. 9.15, the waves on the right are incident on a dielectric boundary, with reflection coefficients r_1 and r_2 (for reflection in regions 1 and 2), and transmission coefficients t_{12} and t_{21} (for transmission from region 1 to 2, and 2 to 1, respectively). Those coefficients are given as

$$\begin{aligned} r_1 &= \frac{n_1 - n_2}{n_1 + n_2} \\ r_2 &= \frac{n_2 - n_1}{n_1 + n_2} \end{aligned} \tag{9.9}$$

and

$$\begin{aligned} t_{12} &= \frac{2n_1}{n_1 + n_2} \\ t_{21} &= \frac{2n_2}{n_1 + n_2} \end{aligned} \tag{9.10}$$

With these definitions, for example, a_r and b_l can be easily written as

$$\begin{aligned} a_r &= t_{12}a_l + r_2 b_r \\ b_l &= t_{21}b_r + r_1 a_l \end{aligned} \tag{9.11}$$

which, after some rearrangement, becomes the transfer matrix for a dielectric reflection, which is

$$\begin{bmatrix} a_r \\ b_r \end{bmatrix} = \begin{bmatrix} 1/t_{12} & r_1/t_{12} \\ r_1/t_{12} & 1/t_{12} \end{bmatrix} \begin{bmatrix} a_l \\ b_1 \end{bmatrix} = M_2 \begin{bmatrix} a_l \\ b_1 \end{bmatrix} \tag{9.12}$$

The power of the transfer matrix method is that it allows us to combine the optical operations (propagation and then reflection) into a single matrix. To represent the relationship between the waves on the right side of Fig. 9.15, in the block labeled n_2, and the waves on the far left side of the first n_1 block, we can multiply the matrices together appropriately. The input to the dielectric is the output from the propagation. The expression

$$\begin{bmatrix} a_r \\ b_r \end{bmatrix} = M_2 M_1 \begin{bmatrix} a_l \\ b_1 \end{bmatrix} \tag{9.13}$$

represents the optical transfer matrix between the waves on the left of the figure and the waves on the right of the figure.

This can be applied to the entire distributed feedback laser structure, with appropriate propagation and dielectric reflection matrixes applied for each of the grating teeth, as shown in Fig. 9.17.

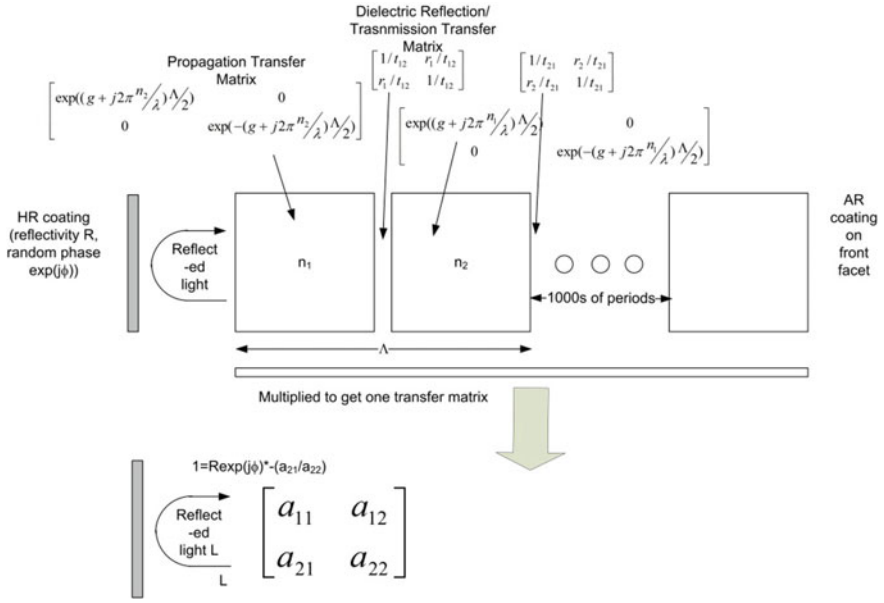


Fig. 9.17 Use of the transfer matrix to model to distributed feedback lasers. The entire operation of a laser is modeled by a single matrix

This single matrix picture is a model of the light propagation inside the structure. One boundary condition is that b_r on the right of the structure is zero (there is no light coming into the structure.) As in Fabry-Perot lasing modes, the condition for single-mode lasing is unity gain and zero net phase. Both these conditions can be concisely expressed as

$$1 = -R\exp(j\phi)\frac{a_{21}(g,\lambda)}{a_{22}(g,\lambda)} \tag{9.14}$$

where the coefficients a_{21} and a_{11} are written explicitly as functions of the gain g and the wavelength λ.

Ignoring phase for the moment (solving Eq. 9.14 for just the amplitude), if the wavelength λ is picked, the necessary lasing gain g can be solved for numerically. Doing this for the relevant range of λ gives the curve $g(\lambda)$ which is the gain envelope curve shown in Figs. 9.7, 9.10, and 9.12.

With phase included, the wavelengths exhibit the same comb of allowed modes that Fabry-Perot laser modes do, and only certain wavelengths of any given structure exhibit the zero net phase that is required for lasing. That gives rise to the points shown in Fig. 9.10. These points lie on the gain envelope, and change of the phase (such as random change of the back facet phase) shifts the allowed wavelengths along the gain envelope curve. With the information about lasing wavelength and gain, anything discussed in the previous sections (gain margin, slope

efficiency, threshold currents, and lasing wavelength) can be calculated. The statistics can be calculated by imposing a random distribution on the back facet phase.

We will leave off the discussion of the transfer matrix method here, except for the extent that we explore it in the examples and problems.

This is a powerful framework to analyze real devices, since variations in length, κ, R, and other parameters can be included. Its major weakness is that it does not simplify the subject particularly. In the next section, we are going to discuss the coupled mode perspective of laser analysis, which is more difficult to apply to realistic devices but does give some insight and another physical picture.

9.6 Coupled Mode Theory

A different way to model semiconductor lasers is through coupled mode theory which we introduce here. This is more analytical than the transfer matrix method (which requires computing power to implement) but is most directly applicable to very simple (anti-reflection/anti-reflection) conditions.

9.6.1 A Graphical Picture of Diffraction

Before we discuss the details of coupled mode theory, let us illustrate a useful way to look at the interaction of light with a periodic structure. Below, we show coherent light incident on a grating structure, and the specular and diffracted orders associated with it.

The usual equation given for the allowed angle θ_m of the diffracted beams is

$$\theta_m = \sin^{-1}\left(\frac{m\lambda}{\Lambda} - \sin\theta_i\right) \tag{9.15}$$

in which the angles are defined in Fig. 9.18, and λ is the wavelength of incident light. Another more graphical picture can be seen in the dispersion-like diagram on the right. This graphical picture will be very helpful in looking at gratings in distributed feedback lasers in the next section.

A graphical way to understand diffraction is to associate a scattering vector $\beta_{\text{scattering}}$ with the grating itself, equal to $2\pi/\Lambda$ (the grating period). This scattering vector $\beta_{\text{scattering}}$ adds or subtracts to the incident light k vector to form the scattered light k vector. The magnitude of the k vector is constrained to be $2\pi/\lambda$, illustrated by the circle on the right. Additions or subtractions to k_x change the diffracted angle as well as the k_x magnitude, but keep the magnitude of k overall the same.

Depending on the shape of the grating, the light may not scatter in all possible multiples of $\beta_{\text{scattering}}$; but that is a detail not relevant here.

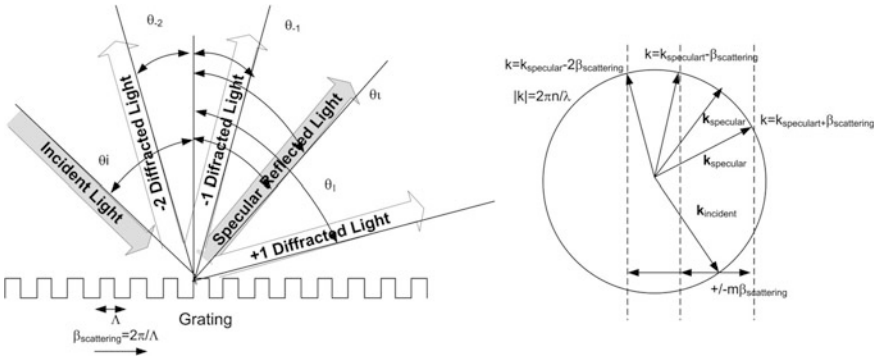


Fig. 9.18 Coherent light incident on a diffraction angle, showing schematically the allowed diffraction directions

Automatically if the scattering vector is too big (and the grating too small, compared to the wavelength of the light), there is no diffraction.

The next section examines what happens in a distributed feedback laser with an included grating.

9.6.2 Coupled Mode Theory in Distributed Feedback Laser

Another perspective on distributed feedback operation is offered by coupled mode theory. Rather than modeling each detailed piece of the distributed feedback structure, in coupled mode theory one steps *way* back and approaches the subject mathematically. That way is perhaps better to get a more intuitive picture of the operation of the device, but it is not quite as applicable a tool to model variations in these devices versus laser parameters. Here, we follow Haus' treatment with the addition of a gain term.[1]

The picture associated with a coupled mode picture is shown in Fig. 9.19. The laser cavity is modeled as medium with gain and a grating, and two optical modes propagate back and forth. Through its periodic scattering of the light wave, the grating continuously reflects one mode back into the other and the forward and backward modes are said to be coupled by the grating.

Though the scattering vector is a vector (in the same way that the propagation constant k is a vector), in this one-dimensional discussion we are going to write these β's as scalars. In some sense, the grated region and the forward and backward modes in it are a 1D diffraction problem. For laser optical feedback, the forward-going mode should be diffracted into the backward-going mode, which should be diffracted again into the forward-going mode. The difference between this and the diffraction diagram of Fig. 9.18 is that in Fig. 9.18, the modes interact,

[1]H. Haus, *Waves and Fields in Optoelectronics*, Prentice Hall, 1984.

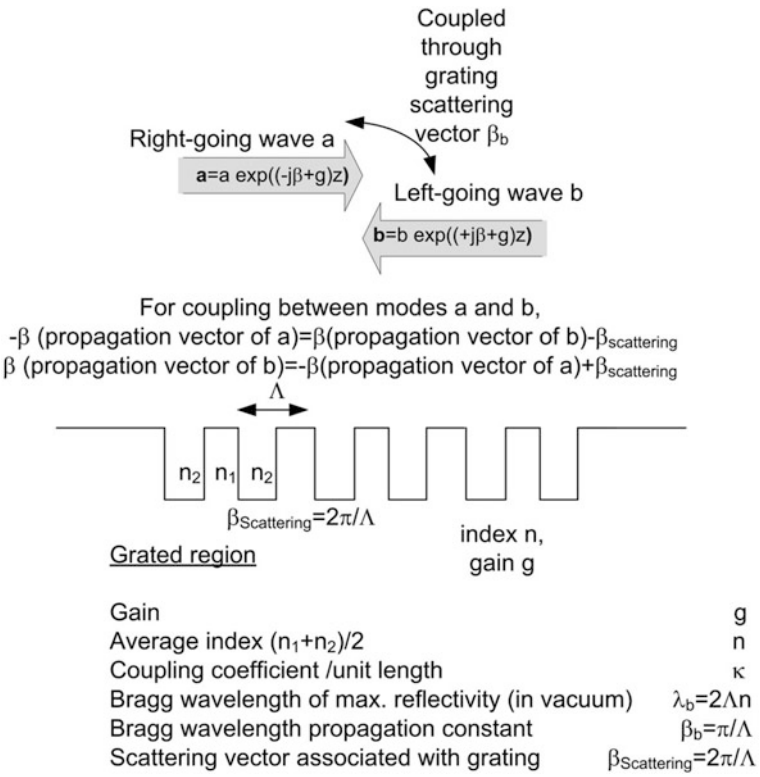

Coupled through grating scattering vector β_b
Right-going wave a
a=a exp((-jβ+g)z)
Left-going wave b
b=b exp((+jβ+g)z)
For coupling between modes a and b,
-β (propagation vector of a)=β(propagation vector of b)-$\beta_{scattering}$
β (propagation vector of b)=-β(propagation vector of a)+$\beta_{scattering}$
Λ
n_2 n_1 n_2
$\beta_{Scattering}=2\pi/\Lambda$
index n, gain g
Grated region
Gain g
Average index $(n_1+n_2)/2$ n
Coupling coefficient /unit length κ
Bragg wavelength of max. reflectivity (in vacuum) $\lambda_b=2\Lambda n$
Bragg wavelength propagation constant $\beta_b=\pi/\Lambda$
Scattering vector associated with grating $\beta_{Scattering}=2\pi/\Lambda$

Fig. 9.19 Two modes coupling in a grated region

diffract, and are gone; here, the condition to confine the modes means the forward and backward modes are continually linked.

For coherent feedback, the forward-going mode ***a*** in the figure above must be precisely coupled into the backward-going mode ***b***, which, when scattered, couples back into the forward mode. The condition for this to happen is if two modes propagate with two propagation vectors, β and $-\beta$, that are coupled together through the grating scattering vector. The relationship between the scattering vector and the forward and backward propagation vectors is

$$\begin{aligned} \beta &= -\beta + \beta_{\text{scattering}} \\ -\beta &= \beta - \beta_{\text{scattering}} \end{aligned} \tag{9.16}$$

Here, let us also identify the Bragg wavelength (which is the wavelength for which the grating has maximum reflectivity) and associated Bragg propagation vector.

$$\begin{aligned}\lambda_{\text{bragg}} &= 2\Lambda n \\ \beta_{\text{bragg}} &= \frac{\pi}{\Lambda}\end{aligned} \tag{9.17}$$

This wavelength is the easiest to picture being coupled by the cavity. The two propagation vectors which are separated by one scattering vector $2\pi/\Lambda$ are +/− the Bragg propagation vector, $\beta_{\text{Bragg}}=\pi/\Lambda$, and so those are the propagation vectors of the forward and backward waves.

For wavelengths different than the Bragg wavelength, the same process occurs. In this case, the propagation vectors become group propagation vectors: These propagation vectors β are associated with the group velocity of the mode and are not necessarily equal to $2\pi/\lambda$. The forward and backward *modes* are then each composed partly of forward- and partly of backward-going *waves* scattered with propagation vectors at the Bragg wavelength.

This process is modeled with a set of coupled equations that describe the change in each optical mode as it propagates. Each mode experiences a phase change (through propagation) and amplitude change (through gain). In addition, a certain fraction of the mode in the opposite direction is coupled into it. The amplitude of that fraction is given by κ, and the exponential terms reflect the change in propagation vector due to scattering.

Mathematically, this is represented as

$$\begin{aligned}\frac{d\boldsymbol{a}}{dz} &= (-j\beta z + g)\boldsymbol{a} + \kappa \boldsymbol{b}\exp\left(-j\beta_{\text{scattering}}z\right) \\ \frac{d\boldsymbol{b}}{dz} &= (j\beta z - g)\boldsymbol{b} + \kappa \boldsymbol{a}\exp\left(+j\beta_{\text{scattering}}z\right)\end{aligned} \tag{9.18}$$

The $\exp(j\beta_{\text{scatterting}}z)$ models the change in the propagation vector of $\boldsymbol{b}$ to couple it back into the $\boldsymbol{a}$ mode.

To make them easier to solve and write, let us make the following two simplifications. First, let us write $\boldsymbol{a}$ and $\boldsymbol{b}$ as

$$\begin{aligned}\boldsymbol{a} &= A(z)\exp\left(-j\beta_{\text{scattering}}z\right) \\ \boldsymbol{b} &= B(z)\exp\left(-j\beta_{\text{scattering}}z\right)\end{aligned} \tag{9.19}$$

This is more than just a mathematical trick. In the range of interest for distributed feedback lasers, the forward-going mode $\boldsymbol{a}$ will generally have a propagation vector close to $-\beta_{\text{bragg}}$. Writing the expression this way means we can neglect the very rapid spatial variation of $\exp(-j\beta_{\text{bragg}}z)$ and instead look at the relative slow change of the envelope function $A(z)$. Substituting Eq. 9.19 into Eq. 9.18 gives us the following set of coupled equations.

$$\begin{aligned}\frac{dA}{dz} &= \left(-j\left(\beta - \beta_{bragg}\right) + g\right)A + \kappa B \\ \frac{dB}{dz} &= \left(j\left(\beta - \beta_{bragg}\right) - g\right)B + \kappa A\end{aligned} \tag{9.20}$$

The expression $\beta - \beta_{Bragg}$ is the difference between the Bragg propagation vector and the mode propagation vector, and is given the symbol δ.

$$\delta = \beta - \beta_{Bragg} \tag{9.21}$$

With that, the equations can be rewritten in a final more concise form.

$$\begin{aligned}\frac{dA}{dz} &= (-j\delta + g)A + \kappa B \\ \frac{dB}{dz} &= (j\delta - g)B + \kappa A\end{aligned} \tag{9.22}$$

These coupled linear differential equations can be easily solved and give a general result of

$$\begin{aligned}A(z) &= A_+ \exp(-Sz) + A_-\exp(Sz) \\ B(z) &= B_+ \exp(-Sz) + B_-\exp(Sz)\end{aligned} \tag{9.23}$$

with a complex propagation constant S equal to

$$S = \sqrt{\kappa^2 + (g - j\delta)^2} \tag{9.24}$$

Let us look at this equation for a little bit and try to see if we can make sense of it. To start, let us assume there is no gain in the structure ($g = 0$). Then, the propagation vector S is

$$S = \sqrt{\kappa^2 - \delta^2} \tag{9.25}$$

The variable δ is the distance from the Bragg wavevector; if the wavelength is the Bragg wavelength, then δ is 0. The further away from the Bragg wavelength we get, the larger δ becomes. If $|\delta|$ is less than $|\kappa|$, then S becomes a real number, and the wavefunctions inside the cavity are decaying exponentials. This is the classical 'stopband' of a Bragg reflector, where wavelengths near the Bragg wavelength decay and do not propagate going into the structure. The amplitude of the stopband is about the same as κ (in appropriate units).

If the gain is nonzero, some positive exponentials can be valid solutions to the propagation equation, and solutions to Eq. 9.25 give the propagation vectors for the envelope functions.

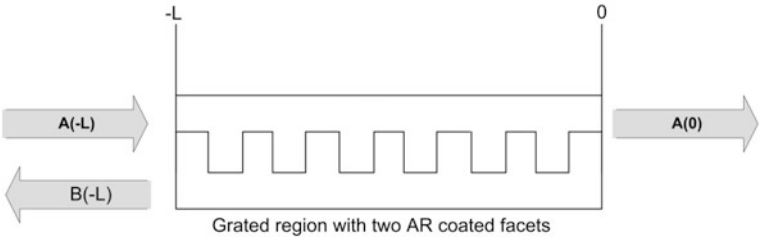


Fig. 9.20 An incident wave, $A(-L)$, incident on a grated region with gain. The reflected wave is $B(-L)$, and the boundary conditions have the wave incident on the structure from the right. The reflection coefficient $B(-L)/A(-L)$ will indicate the wavelengths which support lasing

The ultimate goal is to get some information about g (lasing gain) versus δ (wavelength, written in terms of distance from Bragg wavelength) in terms of κ, device length L, and other factors. To go further in this analysis requires solving the differential equation for some specific conditions. The initial conditions we will look at are shown in Fig. 9.20.

The strategy we will follow is pictured in Fig. 9.20. A distributed feedback cavity of length L, with both facets anti-reflection coated, has light incident on it from the right. We will then find the reflection coefficient, $B(0)/A(0)$. Finally, to deduce the lasing conditions from that, we will find the relationship between δ and g such that, mathematically, there is a reflection without any input. The appropriate boundary conditions are

$$\begin{aligned} A(-L) &= A \\ B(0) &= 0 \end{aligned} \tag{9.26}$$

With these two boundary conditions, the ratio of $B(-L)/A(-L)$ can be found to be

$$\frac{B(-L)}{A(-L)} = \frac{-\sinh(SL)}{\frac{-S}{\kappa}\cosh(SL) + \frac{g-j\delta}{\kappa}\sinh(SL)} \tag{9.27}$$

At the points where the denominator is 0, there can be an output without an input; in other words, there is a lasing cavity. The expression in the denominator defines the relations between gain required and wavelength. It is a transcendental equation with no simple solution, but it can be numerically solved to give the sort of gain envelopes and permitted lasing wavelengths, as shown in Fig. 9.10.

This is a nice mathematical model for a laser which is anti-reflection coated on both sides and, with suitable complex numbers, can accommodate both index-coupled and gain- and loss-coupled lasers. However, it is not quite as straightforward to analyze things like slope efficiency or threshold with asymmetric boundary conditions, and so we take leave of this model except to the extent that it

is covered in the problems. There is no better resource for this topic than the original Kogelnick and Shank paper[2].

9.6.3 Measurement of κ

As we note in the examples of Sect. 9.4, a laser cavity can be designed with a specific period and κ, but what is eventually realized can vary from that. For example, to calculate effective indices requires precise knowledge of the refractive index dependence on wavelength and carrier density (hence laser operating point); typically, these calculations are approximations, refined through an iteration or two of the laser design.

The value of the parameter κ is determined by the fabrication of the device. The designer can control the thickness, composition, and placement of the grating layer to obtain the desired values of n_1 and n_2. Once fabricated, the actual value of the coupling coefficient κ can be estimated by the approximate technique described below.

When there is no gain, there is a region in which light cannot propagate through a grated structure. This region is called the stopband. At very low current densities, there is minimal optical gain in the device, but the spontaneous emission spectra can be easily observed. The stopband shows reduced spontaneous emission in a certain wavelength range. Figure 9.20 shows a measurement of the output spectra at very low current. As shown in Eq. 9.27 with no gain, there is a stopband with reduced emission from the device, and the width of the stopband is related to κ.

In Fig. 9.21, the low output region between the peaks corresponds (roughly) to the stopband between the two peaks of the gain curve. This stopband can be easily measured, and a useful relationship between the measured stopband width and κL is given in the following set of equations.

The parameter κL can be estimated as

$$\begin{aligned} Y &= \frac{\pi}{2}\frac{\Delta\lambda_{\mathrm{sB}}}{\Delta\lambda} \\ kL &= Y - \frac{\pi^2}{4Y} \end{aligned} \tag{9.28}$$

where Y is a parameter, $\Delta\lambda_{\mathrm{sB}}$ is the stopband width, and $\Delta\lambda$ is the Fabry-Perot mode spacing as seen in the figure.

This measurement of stopband and subsequent calculation of κL are tools to analyze the characteristics of fabricated devices and further refine the design. There are also available software tools, such as Laparex (available at http://www.ee.t.u-tokyo.ac.jp/~nakano/lab/research/LAPAREX/, current 11/18) or Glaparex (available through the authors' webpage at Binghamton University). Both these tools

[2]Coupled-Wave Theory of Distributed Feedback Lasers, H. Kogelnick, C. Shank, *J. Applied Physics*, v. 43, pp. 2327, 1972.

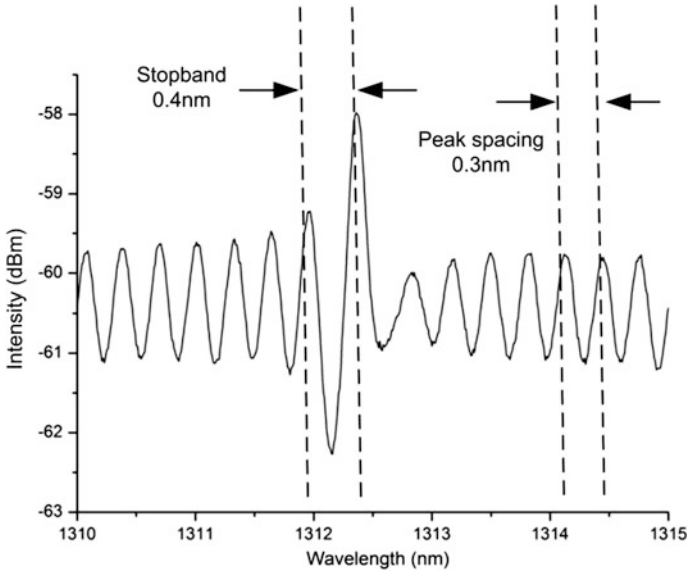


Fig. 9.21 A subthreshold spectra of a distributed feedback laser, showing the stopband, and the spacing between non-lasing modes. Compare this to the threshold above spectra in Fig. 9.1

comprehensively model measured spectra as a function of laser structure and can be used to extract parameters such as κ. Typically, they fit measured spectra and obtain best fit values of κ, reflectivity, and the like.

The value of κL picked for the wafer as a whole determines both the nominal characteristics and the statistics, including the yield of the design to the given specification. It is critically important in achieving a manufacturable and profitable distributed feedback laser design. As we will talk about in Chap. 10, yield is particularly important in the semiconductor business, and a 10% difference in device yield, in products that are approaching commodities, can make a difference between being comfortably profitable and exploring different bankruptcy options.

9.7 Inherently Single-Mode Lasers

One of the things that the reader may note, from Figs. 9.7 and 9.10, is that the distributed feedback lasers we have described so far are only 'mostly' single mode. Because there is a good chance that the gain margin between two lasing modes will be reasonably high, a reasonable number of devices will be single mode. However, the envelope of the gain curve is generally symmetric about the Bragg wavelength and is not by itself, single mode.

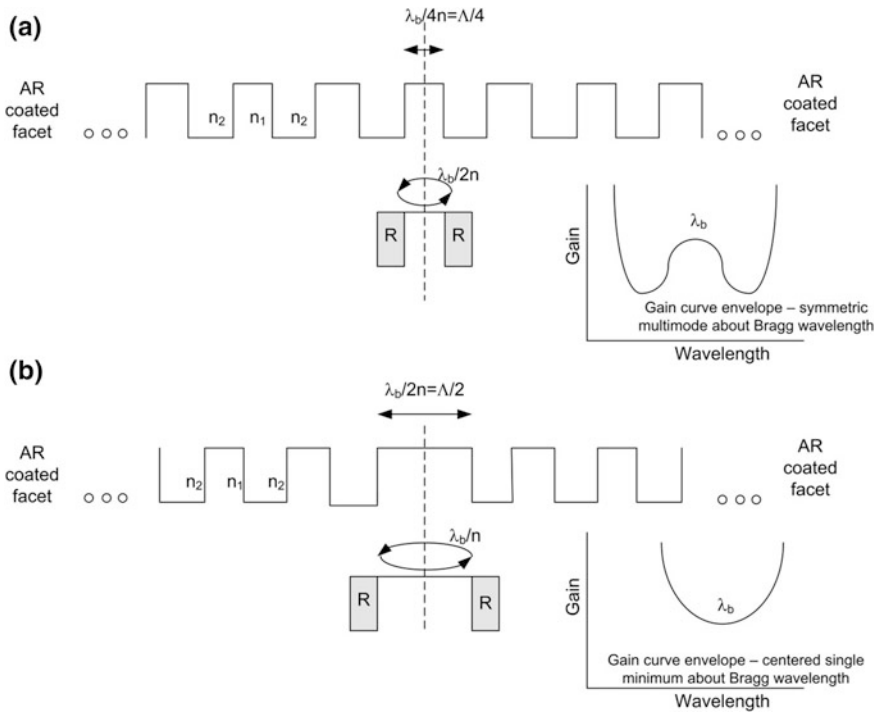


Fig. 9.22 A comparison between a standard laser, with a uniform grating all the way through, and a quarter-wave-shifted device, which has one grating tooth in the center shifted by ¼-wavelength to make the device inherently single mode

A nice picture of why that is so can be seen by considering an ideal AR/AR-coated laser, with the observer located right in the middle of the middle grating tooth, as shown in Fig. 9.22.

Outside of that one grating tooth, the grating goes on, for an equal number of periods on each side. The rest of the grating teeth can be lumped into a single reflectivity *R*. Let us suppose this cavity tries to lase at the Bragg wavelength where the cavity has its point of maximum reflectivity.

Now the observer is in the middle of a very small cavity, watching light bounce from one side, across a ¼-wavelength, to the other side, and back again for another ¼-wavelength. The half-wavelength round trip means that the Bragg wavelength undergoes destructive interference in the cavity, although that is the wavelength who's reflectivity is absolutely the highest.

This problem suggests a solution, shown in Fig. 9.22. Suppose in the very middle of the laser cavity, one grating tooth was widened from ¼λ to ½λ. Considering the observer at the middle of the cavity, the Bragg wavelength goes from

destructive to constructive interference. The fundamental envelope of the gain curve changes from the one on the right to the one on the left. Astonishing, but an extra ¼-wavelength in the material (about 100 nm) can completely shift the characteristics of the device and enable the realization of devices that have close to 100% single-mode yield.

While it is easy to get a uniform grating over an entire wafer using holographic grating techniques, with holographic gratings it is challenging to introduce a single ¼ shift in the center of the device. In addition, the classical argument presented above really holds only for ¼λ-shifted devices with no phase effects from the facets (anti-reflective/anti-reflective coated). For devices with phase effects, like commercial lasers with highly reflective facets, the ¼ shifting technique is very helpful but does not produce perfect single-mode yield.[3]

Recently, it has become commercially feasible to use gratings fabricated with e-beam lithography where every grating tooth is individually written. This technique makes it easy to introduce a ¼λ shift. It is becoming common for commercial devices to have e-beam written gratings with quarter-wave shifts with much higher single-mode yields over the wafer.

9.8 Other Types of Gratings

Figure 9.5 and the coupled mode equations show that for the grating we have considered here, κ is real because the grating is index-coupled. The difference between one periodic material slice and another is just in the refractive index, n.

However, devices which have periodic modulation in gain or loss can also be easily fabricated. If the grating material is absorbing at the lasing wavelength, that will introduce a 'loss grating'; if the grating is actually fabricated to preferentially inject current into the quantum wells, that creates a 'gain grating.' These effects can be mathematically modeled by replacing the real κ in Eq. 9.20 with a positive or negative complex κ for a gain or loss grating, respectively.

The gain and loss gratings can also make the gain envelope asymmetric with respect to the Bragg wavelength, which can be favorable for single-mode yield. Loss gratings of course have some loss associated with them, and so can degrade the threshold or slope. As with almost anything in lasers, it is a trade-off.

[3]In the first edition of the book, the discussion continued, 'this technique is not used typically for commercial lasers.' and went on to explain that it required e-beam grating fabrication, which was not commercially feasible. It now is commercially feasible, and many companies use it. *Proficiscitur in tempore*!

9.9 Learning Points

A. Single-mode lasers are needed for laser communications, both for channel capacity and for long-distance transmission.
B. Since each wavelength can carry different information, many single-mode lasers can carry more information than one multimode laser.
C. Since different wavelengths travel at different velocities, for a good-quality long-distance pulse transmission, the pulse should be composed of a narrow range of wavelengths.
D. There are several methods which can be used to achieve single-mode spacing in lasers.
E. Atomic lasers with very narrow gain regions have inherently single-mode operation; this is not possible in semiconductor lasers, which have broad gain bandwidths of at least 10's of nanometers.
F. Bragg facet coatings or other external wavelength reflectors are also not possible since they do not have a narrow reflectance band.
G. The free spectral range can be made wider than the gain bandwidth by making the lasing cavity narrow. Vertical cavity surface-emitting devices do this and are inherently single longitudinal mode.
H. However, vertical cavity surface-emitting lasers are not good solutions for long-distance fiber communications because vertical cavity lasers have lower slope and lower power output compared to edge-emitting devices.
I. The conventional commercial solution is to include a distributed feedback grating into the laser cavity itself. A long grating with a large number of periods is very wavelength-specific.
J. Though it is similar to a Bragg reflector with a maximum reflectivity at the Bragg wavelength, there are a number of subtle differences. A laser cavity is a mixture of reflector and cavity; wavelengths within the classical stopband of a Bragg reflector can propagate there because there is gain in the cavity.
K. Bragg reflectors (and other optical elements) can be modeled with the transfer matrix method, which allows cascade of many complicated optical elements.
L. Uniformly grated Distributed feedback lasers do not usually lase at the Bragg wavelength of maximum reflectivity, because the reflector is also the laser cavity.
M. A Bragg reflector with no gain has a stopband in which wavelengths are reflected and do not propagate in the cavity. This can be seen by observing spontaneous emission from a laser cavity, in which there is a region of reduced light output.
N. In practical devices that are HR coated on one end and AR coated on the other end, the properties of the laser (including slope efficiency, threshold, and side-mode suppression ratio) vary depending on the exact length of the cavity and the phase of the device when it is reflected from the back facet.

O. Because the properties of these HR/AR devices depend strongly on back facet phase, and back facet phase cannot be controlled since it is defined during the laser cleaving process, the set of devices from a typical identical wafer each have effectively random back facet phase.
P. The yield of a design is determined by the properties of the population; hence, design of a distributed feedback laser should consider the distribution due to random back facet phase as well as the nominal properties.
Q. Quarter-wave shifted devices can have much higher single mode yields than uniformly grated devices and tend to be inherently single mode. They are typically realized with e-beam lithography.
R. Periodic variations in loss or gain (gain gratings or loss gratings) can be used as well as index gratings to realize lasers.

9.10 Questions

Q9.1. Sketch and describe the physical structure and spectral characteristics of the following devices.

(a) Fabry-Perot laser
(b) Lasers with a highly reflective Bragg stack on the front and rear facet
(c) Index-coupled distributed feedback laser
(d) ¼-wave-shifted distributed feedback laser

Q9.2. Would the lasing wavelength of a perfect distributed feedback laser depend on temperature, and if so, how? Compare the temperature dependence of a distributed feedback laser with that of a Fabry-Perot laser. Is there a difference?

Q9.3 If the specifications for a particular laser are SMSR >30 dB and slope efficiency >0.35 W/A, what value of κL should be chosen, based on Figs. 9.14 and 9.15. Estimate the yield to this specification from the best κL.

9.11 Problems

P9.1. Typical values for gain are around 100/cm. Suppose we fabricate an extremely small active cavity device, in which the active region is only 0.1 μm long but the cavity is 3 μm long. (A) What does the value of reflectivity R have to be in order for the gain to not exceed 100/cm in the active region? (B) Assume an absorption of 20/cm. What is the slope efficiency out of the device, in photons out/carriers in? Comment on the general slope characteristics of this device compared to a standard device.

P9.2. We want to design a 300-μm-long distributed feedback laser suitable for a lasing wavelength of 1550 nm, in a material with an index of 3.5. The device should have a negative detuning of 20 nm at room temperature.

(a) What should the gain peak in the quantum wells be (approximately)?
(b) Sketch the output spectra of a fabricated device, along with the output spectra of a Fabry-Perot made with the same material.
(c) Calculate the necessary period for a first-order grating.
(d) Assuming $\Delta n = 0.001$, calculate κ for this material.

P9.3. Consider a grating period twice as big as the Bragg period for a given wavelength.

(a) What is the scattering vector compared to that of a grating at the Bragg wavelength?
(b) Can this grating be used to couple a forward-going and backward-going waves?
(c) Will this wavelength diffract a forward-going wave into any other direction?
(d) What are some potential advantages of this second-order grating?
(e) Suppose the coupling was found to be 12/cm of this geometry (grating thickness, duty spacing, and material). What will the coupling be for this second-order structure?

P9.4. A dielectric stack is designed to be highly reflective at 1550-nm wavelength. If it is composed of two layers, one with an index of 1.5 and one with an index of 2,

(a) Find the appropriate thickness of each material.
(b) Use the transfer matrix method to calculate the reflectivity of a stack of 5, 10, and 25 periods at normal incidence.

P9.5. (a) Implement the algorithm pictured in Fig. 9.16, and use it to calculate the gain envelope for a device with a 200-nm grating period, $\Delta n = 0.005$, $n_{avg} = 3.39$, $R = 0.9$, and a length of 300 μm. Does the calculated Bragg wavelength make sense?
(b) Calculate it for the same parameters but with a length of 200 μm.

P9.6. Show that Eq. 9.11 can be rearranged to give Eq. 9.12.

P9.7 Figure 9.17 shows the interaction of light with a grating. In the process of fabrication of the grating, the grating period is often measured by measuring the diffraction angle of the grating from coherent light. When illuminated by a laser of known wavelength, the diffraction angles unambiguously tell the period Λ of the grating.

(a) If grating has a period of 198 nm, what is the smallest wavelength of light that will diffract?
(b) If light at 400 nm is incident on the grating above at 45°, at what angles will diffraction spots be observed?

P9.8 Download the program Glaparex from the link through the authors' academic Web site.

(a) Fit some of the sample spectra, and run the yield scan to estimate the single-mode yield of the device.
(b) Click the quarter-wave shift button and place the quarter-wave shift about 20–30 μm from the high-reflectivity end of the device. Run the yield scan again and see how it has changed.

Assorted Miscellany: Dispersion, Fabrication, and Reliability

10

"I was wondering what the mouse-trap was for." said Alice. "it isn't very likely there would be any mice on the horse's back."
"Not very likely, perhaps," said the Knight; "but, if they do come, I don't choose to have them running all about."
"You see," he went on after a pause, "it's as well to be provided for everything."
—Lewis Carroll (Charles Lutwidge Dodgson),
Through the Looking-Glass

Abstract

Here we address some topics of importance that don't fit neatly into other chapters. The basic measurement of optical communications quality, the dispersion penalty, is described. We then outline the process flow that takes raw materials to a fabricated and packaged chip. The temperature dependence of laser properties which is particularly important to uncooled lasers is discussed, which leads into the idea of accelerated aging testing for reliability. Finally, some of the failure mechanisms are discussed.

10.1 Introduction

In the previous chapters, we have worked from the theory of lasers, to the theory of semiconductor lasers, to more details about waveguides, high speed performance, and single mode devices. In the process of covering these topics in a systematic way, we have ended up with a complete but basic description of a laser and understanding of its operation.

However, there are many other aspects of laser science, including fabrication, operation, test, and manufacture that should be covered but don't quite fill a whole chapter. In commercial use of these devices, or in research, these areas are less

© Springer Nature Switzerland AG 2020

D. J. Klotzkin, *Introduction to Semiconductor Lasers for Optical Communications*,
https://doi.org/10.1007/978-3-030-24501-6_10

fundamental but are not less important. We want to leave the student conversant with common issues, and as Lewis Carroll says, 'provided for everything', except perhaps horseback-riding rodents.

In this chapter, other aspects of lasers are introduced. Among them are dispersion measurements, typical laser processing flow, differences between Fabry-Perot and ridge waveguide devices, and temperature dependence of laser characteristics.

10.2 Dispersion and Single Mode Devices

In the previous chapter, we described properties of (usually single mode) distributed feedback lasers. As we noted then, one of the motivations for single wavelength lasers is to obtain reduced dispersion; optical signals travel for many km on optical fiber, and because different wavelengths travel at different speeds, a clean set of modulated ones and zeros at the origin can become an ambiguous mess many km later.

Qualitatively that is clear. In this section, we describe more quantitatively how signal quality is evaluated through a dispersion penalty measurement. The basic idea is to measure the *bit error rate* (the fraction of bits that the optical receiver measures incorrectly) as a function of the power on the optical receiver.

The measurement is outlined in Fig. 10.1. Typically in a baseline measurement, a modulated optical signal is coupled to an optical receiver, and a combination of attenuators and amplifiers is used to control the optical power at the receiver end. As the received power is reduced, the number of bits in error increases. A curve typical to the back-to-back curve in Fig. 10.2 is obtained, where the bit error rate goes down as the power at the receiver goes up.

To quantify the effect of dispersion on transmission quality, another measurement is made with a length of fiber in between the transmitter and receiver. Again, amplifiers and attenuators are used to control the power level at the receiver. A second curve of bit error rate versus power level is obtained, this time over fiber.

In real laser systems, increasing optical amplitude is straightforward with erbium-doped fiber amplifiers; however, degradation of transmission quality through dispersion is fundamental. Typically, the power has to be a bit higher (a

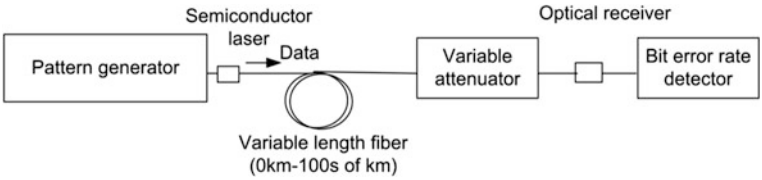


Fig. 10.1 Measurement of dispersion penalty. The signal is put onto a semiconductor laser, through a varying length of fiber (typically $\sim$0 km and the distance over which the dispersion penalty is tested), and then through a receiver and bit error rate detector, which compares the received bit with the bit which was launched. If they disagree, then an error is recorded

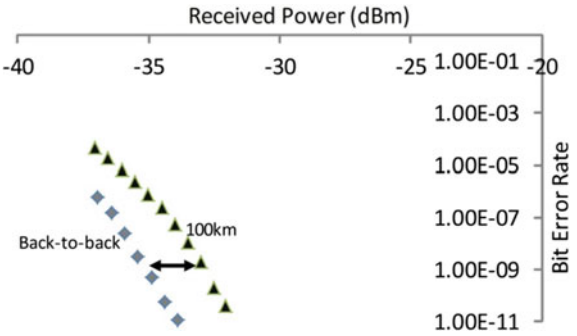


Fig. 10.2 Results of a dispersion penalty measurement. The space between the back-to-back curve and the 100 km curve is the increase in signal power necessary for the data to be transmitted, the dispersion penalty. Typically, it is measured at a specific bit error rate like 10^{-10}

dBm or two) for the error rate to be the same. This required increase in power due to signal degradation from dispersion is called the dispersion penalty.

Typical specifications are 2 dB dispersion penalty over the transmitted signal conditions, for example, 100 km of directly modulated laser signal at 1.55 μm.

As an imperfect analogy, understanding the words to a song on a very soft radio station is easier when there is no static; if there is static, the volume needs to be turned up to understand the words. The dispersion in this case adds the 'static' to the signal.

Since lasers have complicated dynamics, the tests usually done with a pseudo-random bit stream (PRBS) which drive the laser with varying combinations of ones and zeros and ensure that the laser is excited with all possible frequency content.

To aid in connecting dispersion penalty with more fundamental laser parameters, an approximation for the dispersion penalty is given by the expression

$$\mathrm{DP} = 5\log_{10}(1 + 2\pi(BDL\sigma))^2 \tag{10.1}$$

where B is the bit rate (in Gb/s, or 1/ps), L is the fiber length (in km), D is the dispersion of the fiber (in ps/nm km), and σ is the optical linewidth of the signal (Note there are actually many similar expressions used for approximate dispersion penalty. This one is from Miller[1]).

The units for the fiber dispersion penalty D are a bit obscure. It can be read as 'ps' (of delay)/'nm' (optical signal bandwidth)-'km' (of fiber length).

[1]Miller, John, and Ed Friedman. Optical Communications Rules of Thumb. Boston, MA: McGraw-Hill Professional, 2003. p. 325.

Example: A 2.5 Gb/s signal is transmitted using a single mode distributed feedback laser at 1.55 μm over 100 km of standard fiber. This standard fiber has a dispersion of 17 ps/nm km. The dispersion penalty measured as shown in Fig. 10.1 is 1.5 dBm. What is the optical linewidth associated with this transmitter?

Solution: Using Eq. 10.1,

$$\left(10^{1.5/5} - 1\right) \Big/ 2\pi = 0.159$$

$$(0.159)/(100\,\text{km} * 2.5) \times 10^9/\text{s}\ 17 \times 10^{-12}\text{s/nm km} = 0.093\,\text{nm}$$

or about 1.0 Å.

The origin of this 1.0 Å comes from the physics of laser modulation. The wavelength shifts very slightly with the current injection statically (the wavelength of a 'one' is slightly different than the wavelength of a 'zero') resulting in a measurable laser linewidth when modulated. In addition, there is a dynamic chirp during the switch, due to the oscillation of carrier and current density in the core. Because of this, any directly modulated source has numbers of the order of Å.

As an aside, externally modulated sources (like lasers modulated by lithium niobate modulators, or by integrated electroabsorption modulators) do not have this inherent chirp. Because of that, those kinds of directly modulated transmitters can go 600 km or more with appropriate amplification. As another side, the reader is reminded that the dispersion around 1310 nm wavelength in standard fiber is about 0. However, that wavelength is not used for long-distance transmission because the losses are too high (1 dB/km, rather than 0.2 dB/km) and it is more difficult to get in-fiber amplification.

Equation 10.1 also points out how dispersion penalty depends on fiber length, wavelength, and modulation speeds. It is crucially dependent on fiber length because long fibers multiply the difference in propagation velocity between different wavelengths; it is crucially dependent on wavelength because the dispersion penalty depends on differences in speeds at a particular wavelength; and it is crucially dependent on bit rate because slower bit rates require more time for a one to bleed into a zero.

10.3 Temperature Effects on Lasers

A second topic in this miscellaneous chapter is effect of temperature on laser properties. Both the DC and spectral properties do depend strongly on temperature. One additional advantage of distributed feedback devices over Fabry-Perot devices is enhanced temperature stability of the wavelength with temperature changes. To put this in proper

context, fibers can carry many, many channels of information with each channel on a separate wavelength. In order for this work, the wavelength of each channel must be clearly defined and specified so that the various channels do not interfere with each other. As we will see, the temperature affects the operating wavelength of laser devices, but much less in distributed feedback lasers than in Fabry-Perot devices.

For temperature-controlled devices typically used in dense wavelength division multiplexing systems, wavelength control within a nm is maintained by controlling the temperature of the laser source. This is done with an integrated Peltier cooler. For uncooled devices, the inherent wavelength stability of a distributed feedback laser is an advantage.

10.3.1 Temperature Effects on Wavelength

The bandgap of all of these materials depends on the temperature. As the temperature increases, the lattice experiences thermal expansion, and the wave functions of the atoms that overlap to form the band gap change. Hence, the energy band gap becomes smaller and the emission wavelength becomes larger. The typical shift is of the order of 0.5 nm/K. For Fabry-Perot lasers, which lase at the band gap, the lasing wavelength will also change at this rate of 0.5 nm/K.

What about distributed feedback devices, with a fixed grating period? There are slight changes to the period through thermal expansion and to the refractive index through temperature. The net effect is significantly less than that of Fabry-Perot lasers, but is still about 0.1 nm/K.

A third effect is the interaction between lasing wavelength and photoluminescence peak. As discussed in Chap. 9, the difference between the lasing wavelength and peak gain is called the detuning. Typically, the best high speed performance (and the highest differential gain) comes with negative detuning where the lasing wavelength is at lower wavelength than the gain peak. The best DC performance and lowest thresholds are obtained with zero or positive detuning.

Figure 10.3 shows that as the temperature changes, the detuning changes as well. At high temperature, the gain drifts away from the lasing peak, increasing the detuning and the threshold current. At low temperatures, the gain peak approaches the lasing peak and the detuning is reduced. This can change the high speed performance of the device at low temperatures.

10.3.2 Temperature Effects on DC Properties

As the temperature increases, the lasing threshold current increases as well. This happens for several reasons. First, the formula for gain includes the Fermi distribution function for carriers. As the temperature increases, the carriers spread out more in wavelength, and to achieve the same peak gain (set by the optical cavity) more carriers (and hence current) are required. Second, it is the carriers in the quantum wells which contribute to gain. As the temperature increases, a certain

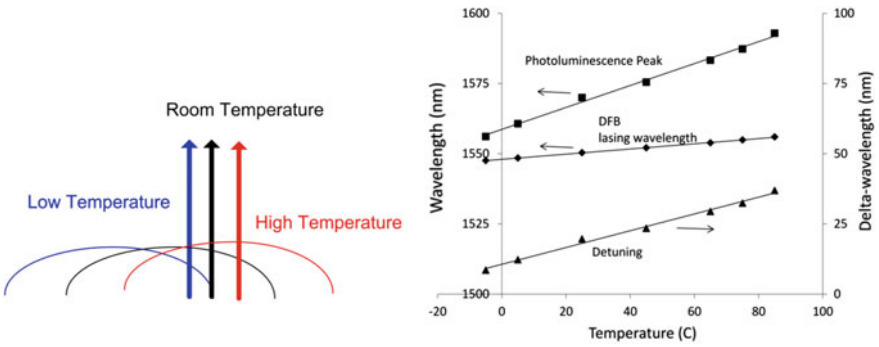


Fig. 10.3 Left, a qualitative illustration of the relationship between gain peak and distributed feedback laser lasing peak versus temperature. The gain peak shifts much more rapidly than the lasing wavelength, creating a detuning that depends on temperature. Right, photoluminescence peak (band gap), distributed feedback lasing peak, and detuning as a function of temperature. The lasing wavelength for a device that is not temperature controlled varies significantly over the operating temperature range

amount of carriers, mostly electrons, escape from the quantum wells and go into the barriers. These carriers do not contribute to optical gain either, and so more current is required to achieve the same peak gain. These mechanisms are illustrated in Fig. 10.4.

The threshold current usually depends exponentially on current, as

$$I = I_0 \exp(T/T_0) \tag{10.2}$$

where T_0 is a constant which depends on material system and, to some degree, on structure. Shown in Fig. 10.5 are two *L-I* curves taken at different temperatures illustrating the change in device characteristics over temperature.

Usually, these DC characteristics are quantified with the T_0 of the device, determined by measuring threshold current versus temperature and finding the T_0 that provides the best fit.

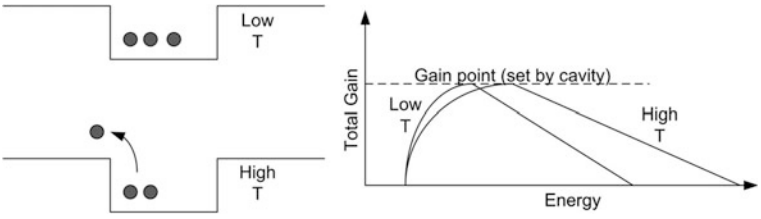


Fig. 10.4 Illustration of the mechanisms for threshold current increase with temperature. Left, carriers escape into the barrier layers; Right, thermal spreading of carriers within the quantum wells. More carriers are needed to achieve the same peak gain

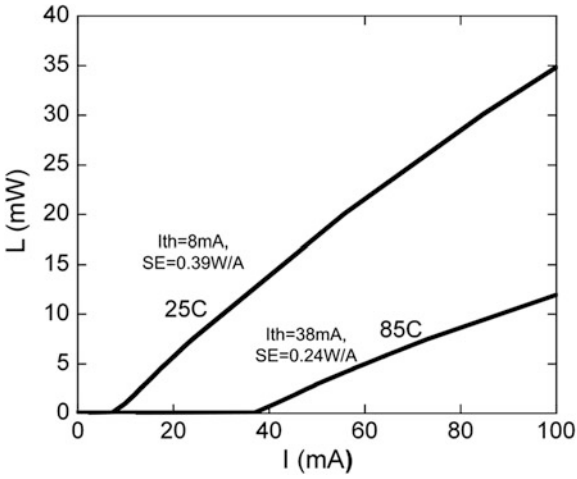


Fig. 10.5 *L-I* curve taken at two different temperatures illustrating the change in laser performance characteristics of the device

Example: In the data shown in Fig. 10.5, find the T_0.

Solution: $I(25\ ^\circ\text{C}) = 8$, $I(85\ ^\circ\text{C}) = 38$, and so

$$\frac{8}{38} = \frac{\exp(25/T_0)}{\exp(85/T_0)} = \exp\left(\frac{25-85}{T_0}\right)$$

or

$$T_0 = (85-25)\Big/\ln\left(\frac{38}{8}\right) = 38\,\text{K}$$

This number of about 40 K is typical of InGaAsP laser systems.

Lasers designed for uncooled use (i.e., without a piezo-electric heater/cooler integrated into the package) must be designed to have reasonable operating characteristics over a broad range of temperature. Typical specifications can be from 0 to 70 °C, or −25 to 85 °C, or more. For those sorts of lasers, T_0 is very important. A high T_0 means device characteristics will vary less with temperature, and a laser with a threshold of 10 mA at room temperature may only be up to 25 mA at 85 °C.

As it happens, the InGaAlAs family of materials (as opposed to the InGaAsP) has a very high T_0, typically 80 K or more; hence, InGaAlAs is the preferred material for high temperature, uncooled devices. The disadvantage to InGaAlAs

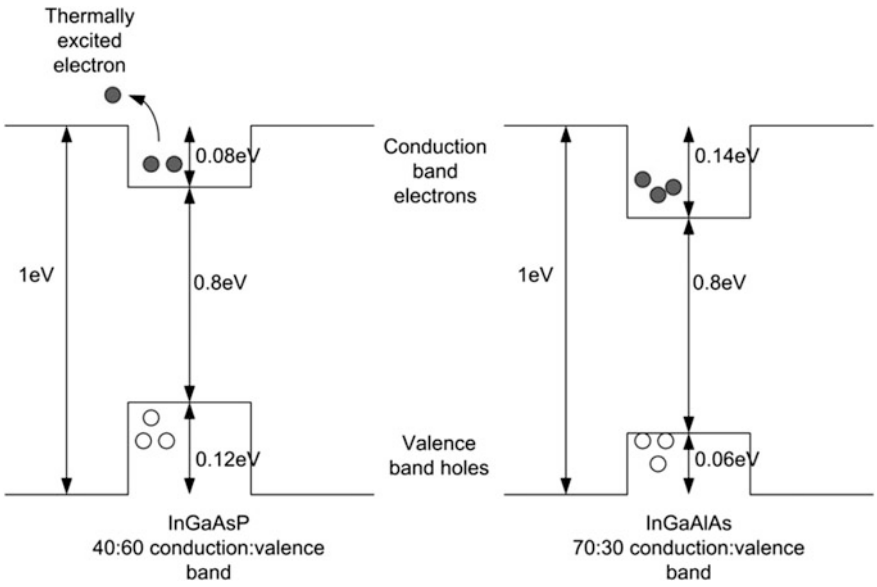


Fig. 10.6 Band structure of InGaAsP and InGaAlAs. The band offsets divide up differently, so that InGaAlAs is much less sensitive to temperature than InGaAsP

(which we will discuss talking about comparison between buried heterostructure and ridge waveguide devices) is that the Al tends to oxidize and so leads to defects in the material. This manifests in better reliability in devices based on InGaAsP compared to devices based on InGaAlAs.

The reason InGaAlAs is better at high temperature is illustrated in Fig. 10.6. In addition to band gap, another important property of laser heterostructures is how the band offset splits up between the valence and conduction bands. For example, a 1.55-μm active region (energy band gap of 0.8 eV) is sandwiched by cladding layers at 1.24 μm (energy band gap of 1 eV). The difference in energy between the core and cladding (0.2 eV) divides up between the valence and conduction band in different ways, depending on the material system.

For example, in InGaAsP material systems, 40% of that 0.2 eV difference appears across the conduction band and the remaining 60% appears in the valence band. The net 'barrier' to electrons is 0.08 eV (not that much different than the 0.026 thermal voltage). Because of that, as the temperature increases, a greater fraction of electrons thermally excite out of the conduction band and into the barriers, and more current is needed to get the carrier density in the wells at the threshold level.

The author whimsically pictures this as a popcorn popper that will lase only when the popcorn is at a fixed level—but the higher the temperature, the more kernels are popped out and wasted. It is a shame to waste popcorn like that!

Luckily, the situation is much more favorable in the InGaAlAs material system. In that system, the barrier breaks up 70% on the conduction band side and only 30%

on the valence band side. The electrons are effectively in a much deeper well and so have much less leakage into the barriers.

In both these cases, it is the electrons who are the important carriers. The effective mass of the electrons is about $0.1m_0$, which is much less than that of the holes, and so they are much more susceptible to thermal leakage.

10.4 Laser Fabrication: Wafer Growth, Wafer Fabrication, Chip Fabrication, and Testing

We have touched upon fabrication in bits and pieces in prior chapters, when it was relevant. Here it is very worthwhile to cover the flow of the laser fabrication process completely in one place. Part of the laser compromises that are made are driven by the materials and processing issues and often it is not the design, but the fabrication issues, which cause problems with laser performance.

In this section, we will first present an overview of substrate wafer fabrication, including the wafer fabrication and the subsequent growth of the active region.

To clarify the terminology, 'wafer growth' means the creation of the wafer, including the substrate and the quantum wells; 'wafer fabrication' means the lithographic processes of making ridges, metal contacts, etc.; chip fabrication is the more mechanical aspects of separating the device into bars and chips and testing it. We also mention (briefly) packaging.

10.4.1 Substrate Wafer Fabrication

All laser fabrication begins with a substrate wafer. This substrate wafer is typically made starting with a seed crystal and a source of the relevant atoms (In and P, or Ga and As) that are exposed to it in molten or vapor form, and then cooling it under controlled conditions in contact with a seed crystal to form a large wafer boule.

A picture of the overall process is shown in Fig. 10.7. In this particular InP wafer fabrication process, a Bridgeman furnace is used to create polycrystalline but stoichiometric crystals of InP. These crystals are then melted together while encapsulated by a layer of molten boric oxide. A seed crystal is then pulled from the melt, and as the layers freeze, a large, single crystal of InP is formed.

The physics of the crystal growth can be quite involved and merit either a detailed discussion, or the merest mention. Here, we stay with the latter and give a qualitative overview.

Once a large single-crystal boule has been fabricated, the wafer flat is marked to show its orientation. It is then cut into thin slices ($\sim$600 μm thick) and polished on one side to form wafers that are ready to be grown. Figure 1.4 in Chap. 1 shows a picture of a typical semiconductor wafer in its ready-to-be-processed state.

Particularly for lasers, the underlying wafer quality is important. Defects in the underlying wafer can eventually make their way to the active region and degrade

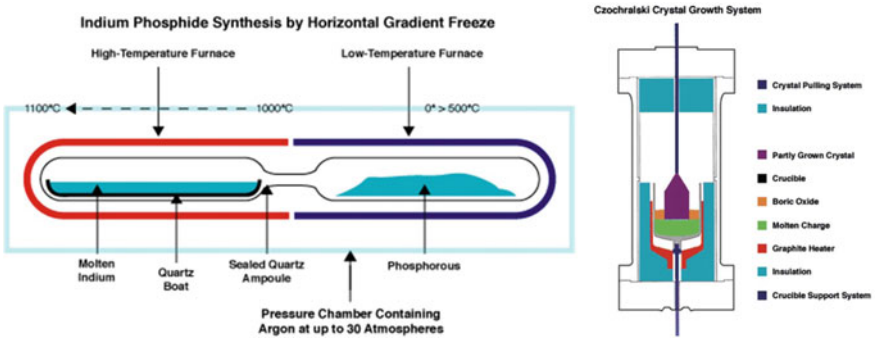


Fig. 10.7 Substrate wafer fabrication. First, In and P are melted and re-frozen in polycrystalline InP; then, the polycrystalline InP is melted again, put in contact with a single crystal seed crystal, and pulled from the metal, to form a large boule which is then sliced into further wafers. Picture credit, Wafer Technology Ltd., used by permission

the device performance. As part of testing, typically a sample of devices is given accelerated aging testing to see how their characteristics change over time. Devices built on wafers with high defect density suffer quicker degradation of their operating characteristics, and it is harder for them to meet the typical lifetime requirements. The idea of reliability testing will be discussed further in Sect. 10.11.

10.4.2 Laser Design

Laser design begins with the detailed specification of the laser heterostructure. The essence of the laser is the active region, which includes the set of layers of quantum wells (which form the active region) and separate confining heterostructures (which form the waveguide). Design of the laser consists of specifying the composition, doping, thickness, and band gap of this set of lasers. A typical laser heterostructure design is shown in Fig. 10.8. Often, in addition to specifying the structure, the required characterization methods are specified as well.

A few comments on the laser structure are made in the diagram.

The top and bottom layers are heavily doped to facilitate contact with metals. The layer below the top layer—which would form the ridge in a ridge waveguide laser—is moderately doped. Most of the resistance in the device is caused by the conduction through this region, and the doping is a trade-off between reduced free carrier absorption and increased resistance.

In this case, the active region of this structure is undoped. This is not always true; often, semiconductor quantum wells are p-doped, which increases the speed but also increases free carrier absorption of the light. The number and dimension of quantum wells are typical of directly modulated communication lasers. This design uses strain compensation, in which the barrier layers (whose only real purpose is to define the quantum wells) have a strain opposite that of the quantum wells, but reduce the net strain (in this case, to zero).

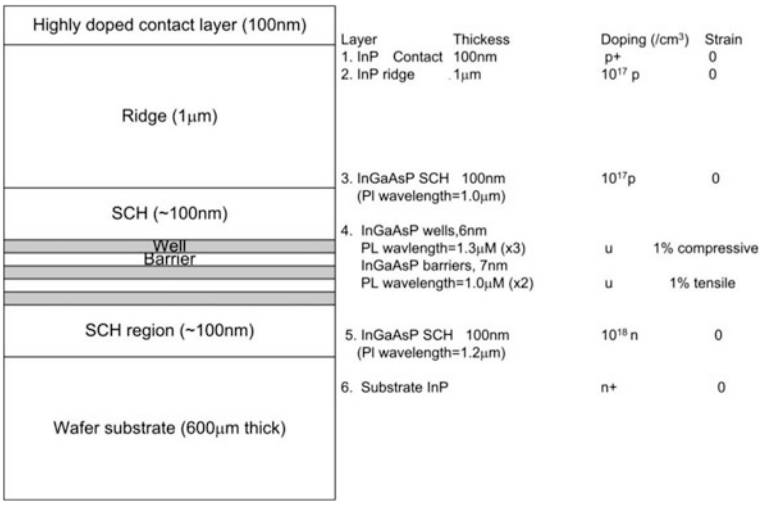

Highly doped contact layer (100nm)
Ridge (1μm)
SCH (~100nm)
Well
Barrier
SCH region (~100nm)
Wafer substrate (600μm thick)
Layer Thickess Doping (/cm3) Strain
1. InP Contact 100nm p+ 0
2. InP ridge 1μm 10^17 p 0
3. InGaAsP SCH 100nm 10^17p 0
(Pl wavelength=1.0μm)
4. InGaAsP wells,6nm
PL wavlength=1.3μM (x3) u 1% compressive
InGaAsP barriers, 7nm
PL wavelength=1.0μM (x2) u 1% tensile
5. InGaAsP SCH 100nm 10^18 n 0
(Pl wavelength=1.2μm)
6. Substrate InP n+ 0

Fig. 10.8 A typical ridge waveguide laser heterostructure design. The doping, thickness, and strain of each laser are specified. Typically, metal contacts are made with the bottom and the top, though some designs have both n and p contacts on the top

10.4.3 Heterostructure Growth

After specification, these layers are fabricated, or 'grown', typically in one of two specialized machines. Either a metallorganic chemical vapor deposition system (MOCVD), or molecular beam epitaxy (MBE) machine, can make layers of the precise thickness, composition, and doping as specified. The basic arrangement of the two techniques is shown in Fig. 10.9 and will be discussed in a little more detail in the subsequent paragraphs. The dynamics and chemistries of the techniques are beyond the scope of the book, and this next section is best appreciated with some microfabrication background.

10.4.3.1 Heterostructure Growth: Molecular Beam Epitaxy (MBE)

An MBE system works by physical deposition. Pure sources of Ga, As, In, or whatever are desired to be grown are independently heated, and the atoms impinge on a source wafer, as shown schematically in Fig. 10.9. They then diffused to an appropriate lattice site and are incorporated into the wafer. The control parameters are typically the temperature of the effusion cells (called Knudson cells) and opening and closing the shutters in front of each cell. The wafer temperature is very important and needs to be precisely controlled.

Typically, the wafer is mounted at the top, and the sources toward the bottom are covered by controllable shutters. To ensure high purity growth of the atoms, the chamber is usually at very high vacuum, and the wafer is transferred in and out through a load lock. Thickness monitoring can be done with an in situ crystal

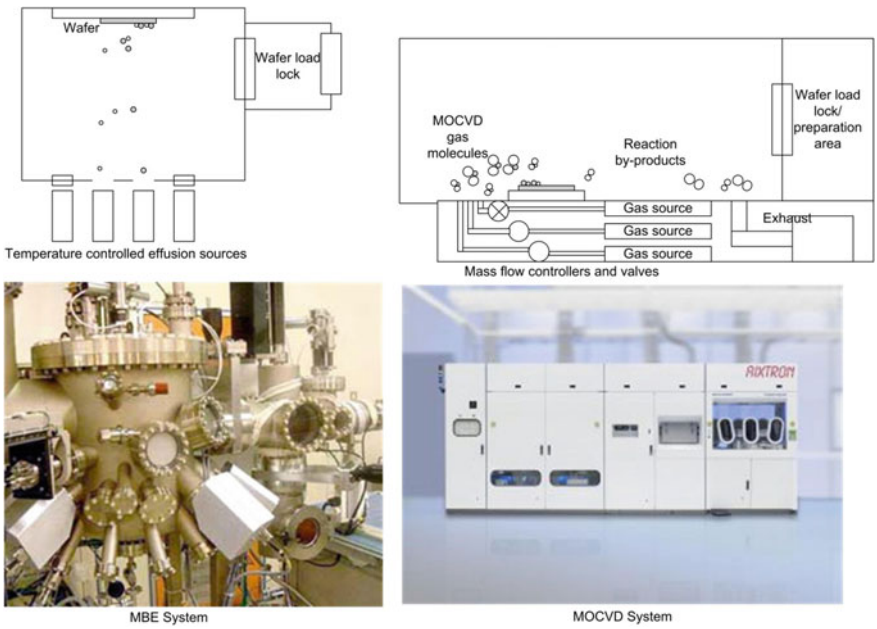


Fig. 10.9 Left, a diagram of an MBE system and a photograph, courtesy Riber; Right, a simple schematic diagram of an MOCVD machine and a photo of an MOCVD machine, courtesy Aixtron. The MBE machine schematically shows atoms being deposited though thermal effusion; in the MOCVD system, chemical reactions occur on the wafer surface and result in the atoms being incorporated into the wafer

thickness monitor, for relatively thick growths. In addition, many MBE machines include a simple electron diffraction system (called reflection high-energy electron diffraction, or RHEED) which can monitor monolayers of growth. The deposition is controlled by the rapid opening and closing of a shutter. Thickness control is more accurate than with MOCVD, and the chemicals used are much safer.

10.4.3.2 Heterostructure Growth: Metallorganic Chemical Vapor Deposition (MOCVD)

In metallorganic chemical vapor deposition (MOCVD), and other vapor deposition techniques, the wafer is loaded into a machine shown in Fig. 10.9. This machine controls the flow rate of various reactive gases (trimethyl gallium, arsine, etc.), and the temperature of the wafer is carefully controlled.

As shown in the figure, as the various gases flow over the heated wafer, they chemically react with it. For example, the Ga atom in trimethyl gallium is incorporated into the lattice of the existing wafer structure, and methane gas is given off as a byproduct. By controlling the flow rate of the gases, and of other gases intending to introduce dopants, the composition and doping density of the wafer can be controlled.

Some of the gases are poisonous or ignite on exposure to oxygen. The MOCVD reactor requires a facility with gas alarms and a charcoal scrubber to cleanse the exhaust. The MOCVD method is almost exclusively used commercially for wafers grown on InP substrates, including devices in the InGaAsP family and in the InGaAlAs-based lasers.

Doing this with accuracy is a very complex task and requires a suite of characterization tools, in addition to the fabrication machine. For example, to grow a p-doped InGaAsP layer (a common laser requirement) requires control of five gases and the wafer conditions. When a wafer recipe is developed, it is usually necessary to measure all of the specified characteristics. Band gap can be measured using photoluminescence; the doping can be measured using Hall effect measurements of conductivity, or sputtered ion microscopy (SIMS); and the strain can be measured with X-ray diffraction. All of these are the beginnings of realizing the thin layer desired.

Wafer growth to some degree is regarded as a 'black art'. Having a body of experience of previously grown similar layers can be enormously helpful.

10.5 Grating Fabrication

At the end of the substrate fabrication and layer growth processes, one is left with a wafer that has the required layers on it and needs to be fabricated into devices with a waveguide, and n- and p-metal contacts. If the device is a Fabry-Perot laser, the layers are the active region, and the wafer will fall into the wafer fabrication diagram pictured in Fig. 10.12. However, if the device is a distributed feedback device with the grating layer below active region, the first step may be patterning the grating layer,[2] followed by an overgrowth of the rest of the devices. Overgrowth means layer growth on a patterned wafer; for distributed feedback lasers (and buried heterostructure lasers, to be described below), overgrowth is necessary. Devices with the grating layer both below and above the active region are commercially used. Below we describe the grating faction steps, followed by the rest of the wafer fabrication.

10.5.1 Grating Fabrication

As discussed in Chap. 9, to realize single mode lasers requires a grating patterned into the device of a particular period. The period is around 200 nm for lasing wavelengths of 1310 nm, and a bit bigger for devices designed around 1550 nm. This is too large to be patterned by simple i-line contact lithography. Most of the

[2]In this example, the grating is under the active region (a common location for it). However, in some processes, the grating is over the active region. In terms of performance, it makes no difference, but one or the other may be more compatible with a given process.

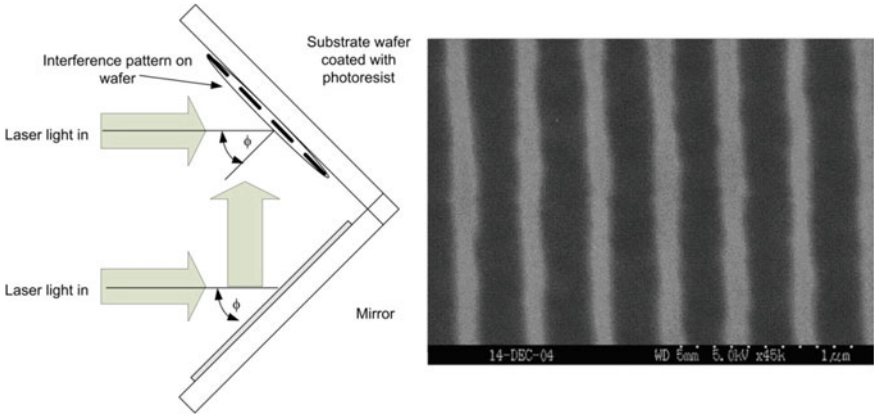


Fig. 10.10 A schematic of a Lloyd's mirror interferometer, in which two interfering laser beams of light form a pattern on the wafer

other steps for lasers are relative large by semiconductor standards and require only 1–2 μm features at minimum.

These gratings are sometimes patterned by holographic interference lithography, as shown in Fig. 10.10. The process goes as follows: A thin layer of resist is spun onto the wafer. A single laser beam, within the range of the resist, is split into two beams and recombined at the wafer surface. The example below is called a Lloyd's mirror interferometer, and with that geometry, the period P of the interference pattern formed is

$$P = \lambda/2\sin(\phi) \tag{10.3}$$

where P is the angle from the normal and λ is the exposing laser wavelength. The minimum achievable period is half the laser wavelength. Wavelengths around 325 nm work well in terms of being within the exposure range of 1800 series photoresist and in producing grating periods down to 200 nm or less.

Then, the wafer is etched, and the resist is removed. What remains is the corrugated pattern on the surface of the wafer.

It is becoming more common for wafers to be patterned by e-beam lithography. Such gratings have tremendous advantages, as they allow every device to have a quarter-wave shift for a much improved single mode yield. In addition, they allow for a greater level of grating design. For example, the quarter-wave shift can be distributed among a larger area, to reduce the peak photon density at the quarter-wave location and reduce spatial hole burning. The duty cycle can also be controlled as way to vary k across the length of the device. Figure 10.11 illustrates some of the potential for these e-beam written gratings.

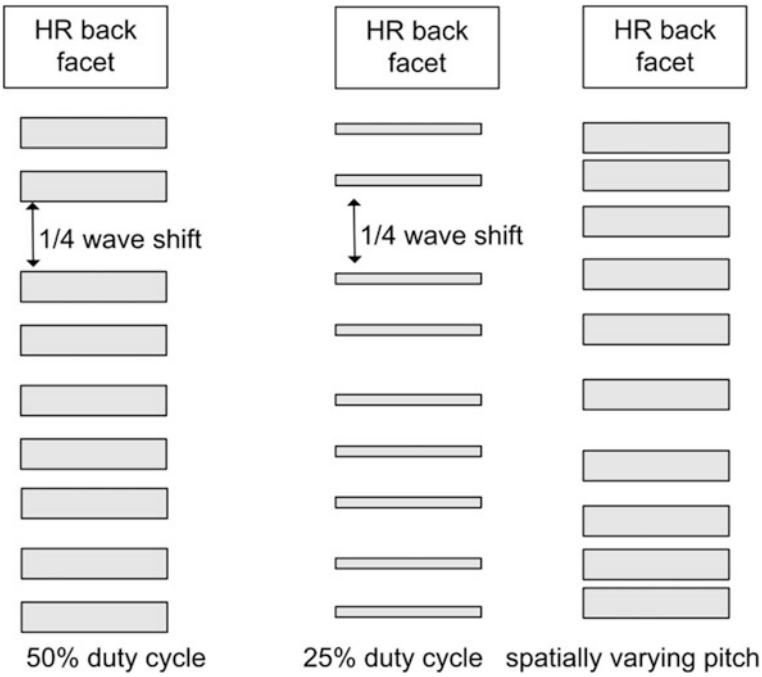


Fig. 10.11 Some features enabled by a e-beam lithography. Every device can be written with a quarter-wave shift. In addition, the duty cycle of the lines can be varied to also control κ, and the pitch can be varied spatially across the devices. All of these features allow control of things like spatial hole burning and single mode yield

10.5.2 Grating Overgrowth

To be effective, the grating has to be integrated as part of the laser heterostructure. The rest of the device structure needs to be grown on top of the grating, while preserving the grating.

This can be challenging; heating up the wafer, as is typically done during wafer growth, causes atoms to move, and diffusion can erode the sharp grating contours. In addition, the overgrowth has to planarize the wafer so the rest of the growths are sharp clean interfaces. Poor overgrowth leads to defects at the growth region and deteriorates the wafer performance. The transition from patterned surface to smooth surface has to happen fairly quickly (within 100 nm or so) as the grating has to be able to affect the optical mode in the device.

Nonetheless, this is largely a solved problem, and the majority of distributed feedback laser are made this way. Figure 10.12 shows an SEM of a grating that has been successfully overgrown. The grating teeth are successfully covered by the rest of the device, and the remaining layers are flat.

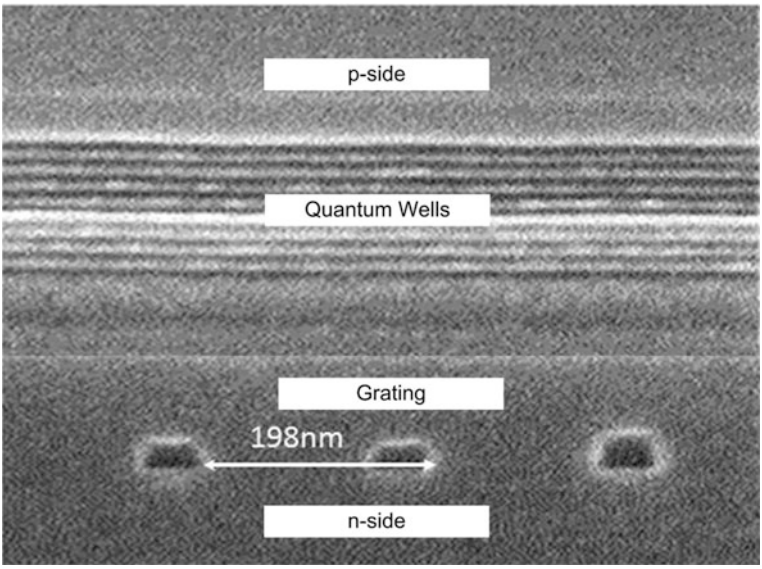


Fig. 10.12 A successful overgrown grating, including quantum wells and surrounding n- and p-regions

10.6 Wafer Fabrication

In this section, we will illustrate the process of turning a wafer (including the substrate, and the initial grown layers) into laser devices. Here the simplest practical device, a ridge waveguide, is shown first, and variations on that basic process are shown for distributed feedback devices and buried heterostructure devices. The latter two incorporate overgrowth which significantly complicates the process.

10.6.1 Wafer Fabrication: Ridge Waveguide

For Fabry-Perot ridge waveguide devices, fabrication starts here immediately after heterostructure growth, and the entire active structure can be grown in a single growth. For distributed feedback lasers, fabrication continues here after the grating layer has been grown, the wafer removed and patterned, and the rest of the heterostructure then overgrown on the patterned grating.

Figure 10.13 shows the fabrication flow of a simple ridge waveguide device. Additional steps which are necessary for buried heterostructure devices will be illustrated in Sect. 10.4.2. The first two steps shown below (grating fabrication) are necessary for distributed feedback devices only (Fig. 10.13).

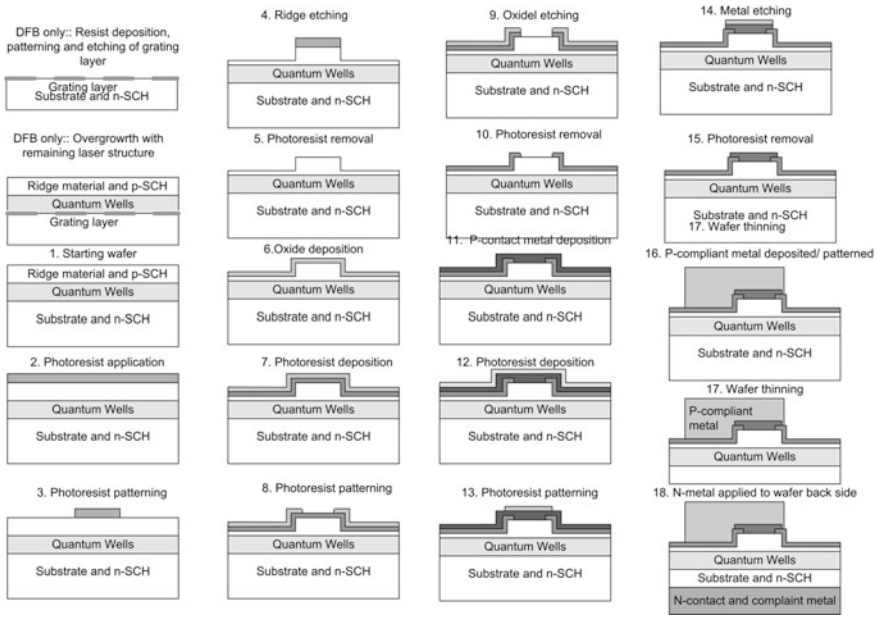


Fig. 10.13 A simple fabrication process overview for a ridge waveguide laser

The first two steps are only for distributed feedback lasers. These steps involve patterning the grating layers and then overgrowing the rest of the structure. For Fabry-Perot devices without a grating, wafer processing starts with the wafer layers already grown on the step labeled 1. A typical first step is etching the ridge (shown in steps #1–5). The ridge etch can be just a wet chemical etch with only a photoresist mask, or (more typically) involve intermediate steps of depositing masking layers of oxide or nitride, patterning them with photoresist, and then using the oxide as a mask for a dry etch. Dry etching has the advantage of making a more vertical sidewall and being more controllable.

The next step is depositing some sort of dielectric insulation on the wafer, so the metal layers to be deposited will not make electrical contact to the wafer except on the ridge (steps #6–10). Then, contact metal is deposited and etched (steps #11–15), leaving p-metal with an ohmic contact on the top of the p-ridge. Finally, a compliant metal pad (typically much larger and thicker) is deposited on top of the contact metal, to allow a place to make external electrical. Typically, the compliant metal is Au (the resist deposition-pattern-develop-metal etch- resist remove steps are omitted, as they are quite similar to the sequence for contact metal).

The wafer is then lapped, which means it is ground down to about 100 μm thickness. Typically, this is done by fastening the front surface of the wafer to a puck with wax, and grinding off the back surface until the thickness is as desired. Thinning the wafer is required in order to be able to divide into reasonably sized bars later.

The n-contact and compliant metals are then applied to the n-side. The wafer is then annealed, to make good ohmic contact to the wafer.

There are additional steps which can be done. For example, sometimes the metal on the n-side is then patterned, which requires an two-side alignment between the metal on the back side and the metal on the front side, as well as the same metal-deposition resist-deposition pattern etch remove cycle as shown in Fig. 10.12 for the p-contact and p-compliant metal.

10.6.2 Wafer Fabrication: Buried Heterostructure Versus Ridge Waveguide

This book has been focused on lasers in general, but here we would like to focus on the two common single mode laser structures—buried heterostructures and ridge waveguide devices—the specific issues associated with both, and the particular differences in fabrication.

Figure 10.14 on the left shows a buried heterostructure device on a 10-μm scale. The heart of the device (the active region) is the small rectangle indicated by the arrow. That is where the quantum wells and the grating layer lies. The filler around it is InP (typically in an InGaAsP system) that serves to funnel the current injected in the top into the relatively small active region. In this structure, the active region is physically carved from the pieces around it.

The picture on the right show a completed ridge waveguide device. The ridge waveguide device is much simpler to fabricate than a buried heterostructure device. The basic fabrication consists of just a simple ridge etch, and the various etches, dielectric deposition, and metallization.

The extra processes for buried heterostructures are shown in Fig. 10.15. Typically, the first step is etching away the mesa, often with a wet etch. Wet chemical etching is thought to form a better, more defect-free surface for overgrowth than a dry etch. The wafer is then put back into a metallorganic chemical vapor deposition,

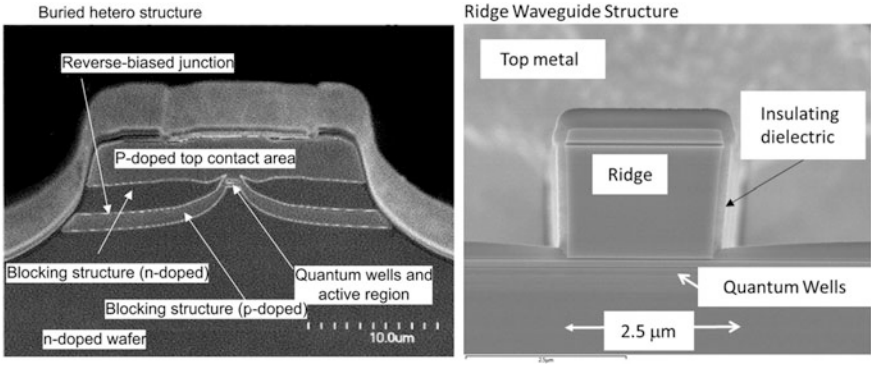


Fig. 10.14 Left, a buried heterostructure laser; right, a ridge waveguide laser

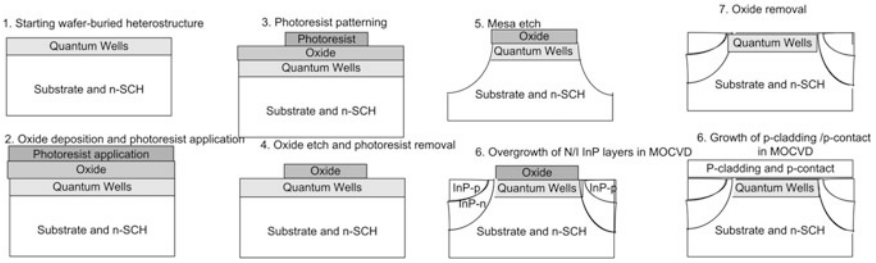


Fig. 10.15 Fabrication process for buried heterostructure wafers

and the active region is overgrown. The process of this overgrowth serves to planarize the wafer again, so that subsequent processes, like dielectric deposition, metal deposition and patterning, can be done on a flat wafer.

It is the doping in the overgrowth that makes these overgrown layers into blocking layers. Typically, these blocking layers are grown either undoped (i) (which has very low conductivity compared to the doped contact layers) or grown (from mesa upward) with a p-doped layer followed by an n-doped layer. On top of that (now top) n-doped layer, the p-cladding layer of the laser is grown. When that layer is positively biased, the junction indicated on the figure is reverse biased, and little current can flow through it. The 10-μm-wide region at the top of the structure shown can be biased, but current will still be funneled only through the active region.

There are advantages and disadvantages to such a structure which are shown in Table 10.1 and discussed below.

Buried heterostructure devices are certainly more complicated to fabricate. In particular, these blocking layers have to be overgrown, which means the fabricated wafer with mesas on it needs to be put back into the MOCVD and have new layers grown upon it. The growth process has to be given low defect densities, or the laser performance and reliability will suffer. In addition, this sort of blocking structure often has reverse bias capacitance associated with the blocking layers, and as discussed in previous chapters, this capacitance, along with residual resistance, can impair the high speed performance.

Table 10.1 Advantages and disadvantages to ridge waveguide and buried heterostructure devices

Laser type	Advantages	Disadvantages
Ridge waveguide	Easy to fabricate—no overgrowth	Lower current confinement Lower optical confinement Generally worse DC *L-I* performance
Buried heterostructure	Better current confinement Better optical confinement Overall better performance	Overgrowth required Parasitic capacitance associated with blocking layers Less reliable overall than ridge waveguides

Additionally, it is difficult to get high-quality overgrowth of Al-containing materials.

The advantages are the structure which does an excellent job of isolating the current, and confining the light, to only the active region. Buried heterostructure devices tend to be the highest performance devices in terms of slope efficiency and threshold current.

The ridge waveguide structure shown on the right of Fig. 10.14 is a much simpler structure. As discussed in Chap. 7, the waveguide is formed by the ridge over a section of the active region. The optical mode sees a bit of the ridge, and so the effective index of the optical mode is a bit higher under the ridge.

Fabrication is very simple, as illustrated in Fig. 10.13. The ridge is just etched down to just above the active region (etching through the active region, leaving an exposed surface and unterminated bonds, effectively introduces defects into the active region.) Typically, an insulating layer like oxide is put down around the ridge, and a hole is opened at the top of the ridge, exposing the contact layer, to which metal contact can be made.

The current is then injected through the top p-cladding ridge directly into the active region.

The trade-off for this straightforward fabrication process is that optical (and current) confinement is not as good as with buried heterostructures, and often slope and threshold are not as good.

10.6.3 Wafer Fabrication: Vertical Cavity Surface-Emitting Lasers (VCSELS)

As long as we are discussing different common types of lasers, we had best briefly mention the fabrication of vertical cavity surface-emitting lasers (or VCSELs), as pictured in Fig. 10.16. Though they do not have a huge place in high-performance telecommunication devices today, they do have significant advantages in both fabrication and testing, and so it is appropriate to at least briefly describe them. At some point, their natural disadvantages may be overcome, and they may become the technology of choice.

Unlike the devices we have discussed before, VCSELs emit light in a vertical direction normal to the wafer. The mirror is formed by Bragg stacks above and below the active region.

To produce these structures on a GaAs substrate, first, alternating layers of GaAs and AlAs are grown on the wafer through MBE or MOCVD. In this case, the layers are grown to form a Bragg mirror (similar to what is shown in Fig. 9.5). AlAs and GaAs have significantly different refractive indices, but remarkably, almost the same lattice constant: Therefore, many pairs of layers can be grown one after another to form a high reflectance bottom mirror, without creating dislocations.

Then, a thin active region of a few quantum wells is grown. Typically, the quantum well region is centered in the optical center of the cavity. Another set of p-doped GaAs/AlAs layers are grown on top of that region, and a round circular

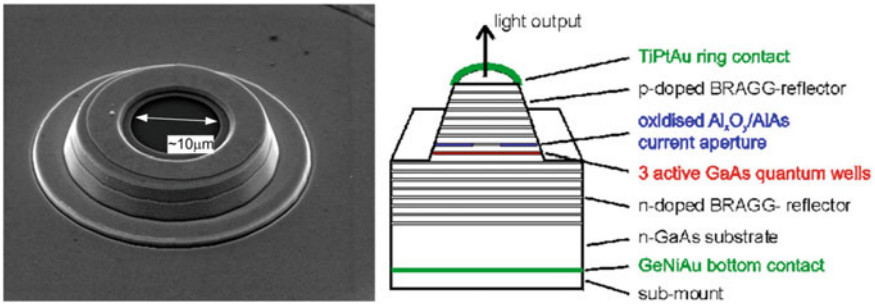


Fig. 10.16 Left, view of a VCSEL mesa. The light is emitted out of the top and bottom. Right, a schematic picture of a VCSEL. The mirrors are provided by many pairs of Bragg reflectors. From *Journal of Optics B*, v. 2, p. 517, https://doi.org/10.1088/1464-4266/2/4/310, used by permission

region is etched to define the lasers in a region a few microns in diameter. Typically, a metal contact is put in a ring around the top of the device. Often an oxide current aperture is formed in the top mirror stack by oxidizing the exposed AlAs layers (making them non-conductive) so as to funnel current only to the center of the device. The edges of the top Bragg stack are nicely exposed after the mesa etch, and the usual tendency of Al-containing compounds to oxidize (thus making it difficult to make reliable buried heterostructure Al-containing devices) is used to advantage, by intentionally oxidizing Al to make it not conductive.

The advantages and disadvantages of VCSELs are shown in Table 10.2. Fundamentally, the advantages are that many more devices can be fabricated on a wafer; they are intrinsically single lateral mode because the optical cavity is so short; and, their far fields are inherently low divergence and couple nicely to an

Table 10.2 Advantages and disadvantages of vertical cavity surface-emitting lasers compared to edge-emitting lasers

Laser Type	Advantages	Disadvantages
Edge-emitting lasers (both ridge waveguide and buried heterostructures)	Overall higher performance-slope, temperature	Generally have to separate before testing Much bigger—fewer devices per wafer
Vertical cavity surface-emitting lasers	Easy on-wafer testing Naturally single lateral mode Excellent far field for coupling to fiber	Limited generally to GaAs-based substrates (due to natural AlAs/GaAs mirror system)[a] and wavelengths <880 nm Generally poor performance over temperature Generally lower power output

[a]Many, many different versions of InP-based VCSELs have been realized in research laboratories. However, as yet (2012) they do not have a market presence in long wavelength telecommunication lasers

optical fiber. Their disadvantages are worse DC performance, as well as the very major disadvantage, for telecommunications use, that there really are no natural mirrors that match well to InP substrates.

10.7 Chip Fabrication

After the lasers have been fabricated, there are many more mechanical steps necessary to turn this wafer, with thousands of devices on it, into thousands of mechanically separated individual devices. The basic flowchart, starting with a fabricated wafer, is shown in Fig. 10.17 (This is a typical process for edge-emitting devices. Processes with surface-emitting devices like VCSELs are very different.).

As we will see in Sect. 10.8, there are substantial advantages in testing things as soon as possible. Labor invested in bad chips has a cost. Hence, being able to quickly identify that the wafer as a whole is below specification is advantageous. If the wafer has to be fabricated into many chips that are individually tested and then discovered to be below specification, then time (which is money) has been invested into bad, unsalable product which has to drive up the cost of all of the remaining devices which fall within the specification. Most companies find some way to do some form of on-wafer testing.

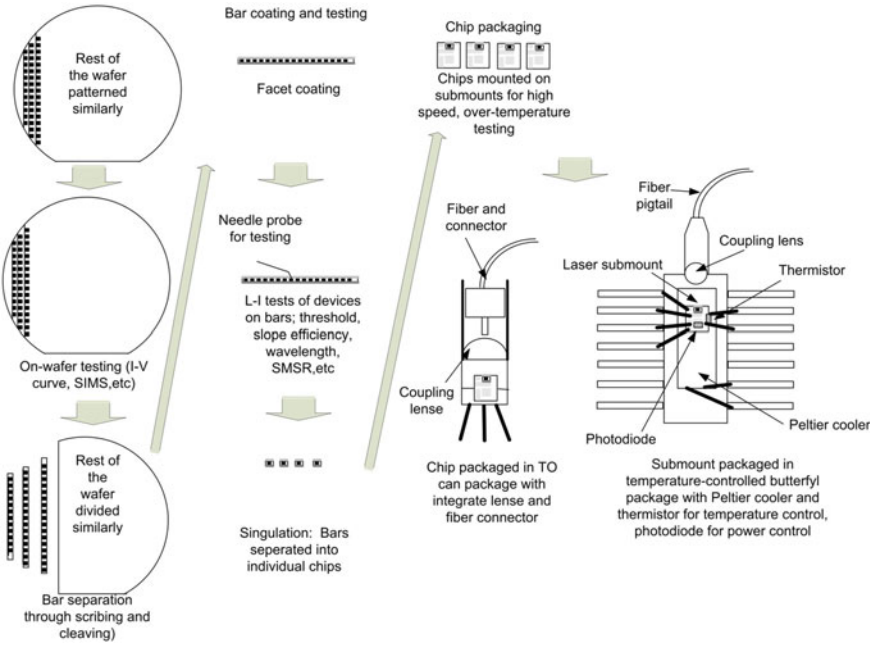

Rest of the wafer patterned similarly
On-wafer testing (I-V curve, SIMS,etc)
Rest of the wafer divided similarly
Bar separation through scribing and cleaving)
Bar coating and testing
Facet coating
Needle probe for testing
L-I tests of devices on bars; threshold, slope efficiency, wavelength, SMSR,etc
Singulation: Bars seperated into individual chips
Chip packaging
Chips mounted on submounts for high speed, over-temperature testing
Fiber and connector
Coupling lense
Chip packaged in TO can package with integrate lense and fiber connector
Fiber pigtail
Coupling lens
Laser submount
Thermistor
Peltier cooler
Photodiode
Submount packaged in temperature-controlled butterfyl package with Peltier cooler and thermistor for temperature control, photodiode for power control

Fig. 10.17 Chip fabrication flow, from fabricated wafer to packaged chip. See text for discussion of various points on the process

This may be as simple as testing the electrical (metal) connections or the *I-V* part of the LIV curves ranging up to nearly full device performance tests.

After the wafer test results are done, if the results merit it, the wafer is typically divided into bars. These bars are cut out of the wafer through the process of scribing and cleaving. First, a small scratch is made on the wafer surface, parallel to one of the wafer planes. Then, the wafer is snapped along the scratch line, cleaving along one of the crystal planes. In Fig. 5.9, the scribed (rough) and cleaved (clean) areas can be clearly identified. The cleave is very important to form an optical quality facet on the edge of the device. For the bar to cleave properly, the optical cavity has to parallel to one of the wafer planes.

The necessity to cleave is one reason the wafer must be lapped (thinned) down to about 100 μm. In order to get 200–300-μm-wide bars reliably, the wafer should be about as thick as the bar width. In addition, the thin wafer aids in the heat removal from the device. InP (and GaAs) has much poorer thermal conductivity than the metal layers that will be put on top of them.

The bars are then facet-coated: A layer or layers of some dielectric material are put on the facet to either reduce or enhance the reflectivity and engineer the emission from the device. For a distributed feedback laser, this facet coating has the purpose of killing the Fabry-Perot modes, so the only optical feedback is the wavelength-sensitive grating feedback. For a Fabry-Perot device, the coatings engineer the emission so that most of the light going out comes out the front end and is coupled to the fiber. A modest amount of light ($\sim 15\%$ typically) is coupled out the back and used to monitor the amount of light from the front facet in situ.

After facet coating, the bars can be tested again. At this stage, things like side-mode suppression ratio (SMSR) and threshold current can be reliably tested. The passing chips on the bar are usually packaged onto a sub-mount, which is a small piece of alumina or aluminum nitride with metal traces on it. The sub-mount often has provision for mounting a back-facet-monitoring photodiode. Once mounted on sub-mounts, high speed tests can be done; however, since it is not yet hermetically sealed, very low temperature tests are not possible due to condensation of water onto the cooled facet.

Finally, sub-mounts passing that test are packaged into device packages, shown at the end of Fig. 10.16. Then the devices can be given full performance testing, including over temperature.

Some performance parameters of the lasers (such as side-mode suppression ratio) are tested on every fabricated device, as they can vary a lot from device to device from the same wafer. Other performance parameters (e.g., bandwidth) are ‘guaranteed by design’ and are tested only on a sample basis.

10.8 Wafer Testing and Yield

After the laser chip is fabricated and before it is sold to a customer, it needs to be tested. Semiconductor laser yields are nothing like integrated circuit yields, and every single device needs to be tested, to verify that it meets all the product specifications.

For a successful commercial operation, laser testing is very important. Unlike strictly electronic devices, semiconductor lasers vary significantly from device to device. Some of this variation is fundamental (e.g., from random facet phase in distributed feedback devices), and some of it is simply due to the extreme sensitivity of these optical devices to material quality.

A successful company that is trying to manufacture devices needs to reduce the costs as low as possible, and one of the ways to do that is through intelligent testing.

Testing devices (particularly, packaging for testing) does cost money. It is beneficial to find bad chips as early as possible before they have been packaged. As an extension to that idea is that if it is possible to test things on a wafer, do that and avoid the labor of cleaving off bars and testing them, or mounting chips on sub-mounts to test them. The point is that testing does both cost money and time, and testing capacity can also be a bottleneck for the number of chips produced.

One simple useful concept here is the idea of yielded cost: How much does a good wafer or laser chip cost? The yielded cost Y.C. is defined as the cost C of the operation divided by the yield of the operation, as

$$\text{Y.C.} = C/\text{yield} \tag{10.4}$$

For example, if it costs \$10 to package a laser in a TO can, and the yield when tested to the TO can specification is 80% (0.8), then the yielded cost per good device is \$12.50. To make 80 good devices, you will have to package 100 at a cost of \$10/each, and so, it will cost \$12.50/each per every good one. If the yield can be reduced on the per/wafer steps to be increased on the/chip step, it is almost always a worthwhile trade-off.

An example of this sort of optimized testing is illustrated in Table 10.3. The numbers in the table may be outdated, but the idea is clear. If a bad wafer can be identified early and discarded, the cost of chips eventually produced is reduced.

In the first method, every wafer is divided into chips, and every chip is tested, while in method B, wafers which are projected through some means to have a lower yield (perhaps their contact resistance is higher) are simply discarded. Here, just throwing out wafers which would have a lower yield and making another wafer lowers the cost of each final package by 10%.

In addition, there are often opportunities to eliminate expensive tests (like, tests over temperature) in favor of finding correlations (like room temperature measurements).

Table 10.3 Illustration of two different strategies of laser testing

Step	Method A		Method B	
	Cost	Yielded cost	Cost	Yielded cost
Wafer fab + test	$5000 (100%)	$5000/wafer	$6000 (80%)	$7500/wafer
Device fabrication	$30 (80%)	$37.50/chip	$30 (80%)	$37.50/chip
Device test	$50 (28%)	$178.57/chip	$50 (35%)	$142.85/chip
Total yielded cost/chip		$216		$180

Method A does not do on-wafer testing, and so has a slightly lower average yield than method B, which does on-wafer testing and eliminates 20% of the wafers but results in a higher chip test yield

10.9 Reliability

In addition to performance tests, like threshold, slope efficiency, side-mode suppression ratio, and the like, semiconductor lasers must have a certain reliability in order to be sold commercially. This means that they are at least expected to perform within specifications for some given lifetime. Guaranteeing this (or at least, assuring the likelihood of it) is a major effort and part of the quality that goes into semiconductor devices.

In this section, we briefly describe the process by which laser reliability is quantified. To illustrate the idea, we will walk the reader through an analysis of laser reliability, though the specifics of the procedures followed vary company to company.

10.9.1 Individual Device Testing and Failure Modes

It is impossible of course to directly test whether a laser will last for 10 or 25 years or any reasonable nominal lifetime. To indirectly test this, laser companies typically do accelerated aging tests, in which devices are operated continuously at levels well above its normal operating characteristic. For example, a sample of lasers intended for cooled use at around 25 °C might be tested at 85 °C. The devices are kept at 85 °C for months and months and during that time, the current required for fixed power output, or the power output for fixed current, is monitored. Additionally, device aging can be accelerated by operating at currents significantly above specified operating currents.

Since there is substantial variation from device to device, typically a few samples of a particular device are used, and aging rates from each device in a sample are computed.

Lasers have different failure modes. Most of the devices in Fig. 10.18 are shown experiencing *wearout failure*, which is a gradual performance degradation attributed to the accumulation of defects in the active region. This manifests as an increase in the current required for fixed power, or a decrease in the power output for fixed current over thousands of hours. The rate of degradation can be modeled as a %/khr.

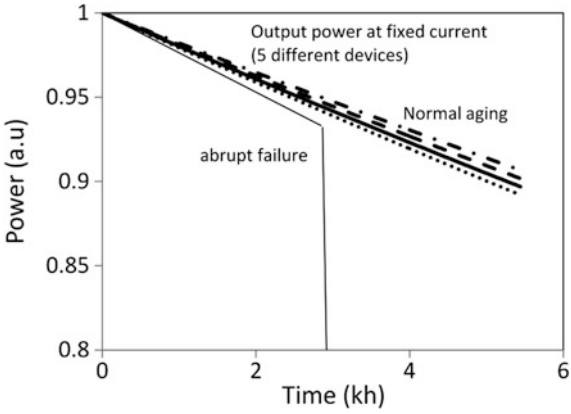


Fig. 10.18 Aging data from a sample of lasers. Aging conditions are typically much harder than operation conditions and are extrapolated down to operating conditions to predict reliability there

Also shown in Fig. 10.17 is an example *of random failure*. In these failures, the laser very suddenly fails by a mechanism not due to gradual defect accumulation.

The root cause of these failures is defects, which are often created at regrown interfaces (like the blocking layers in a buried heterostructure devices). The growth of these defects and encroachment into the active region as dislocations can often be observed in transmission electron microscopy (TEM) measurements after failure.

Figure 10.19 shows some TEM images taken through the plane of the grating in a buried heterostructure device. Dislocation loops are seen going from the regrown interface to the active region of the device. This is one of the weaknesses of buried heterostructure devices, which are less reliable than ridge waveguide devices because of the proximity of the blocking layer regrown interface to the active region.

Sometimes devices suddenly fail due to damage to the facet from catastrophic optical damage. With catastrophic optical damage, the facet absorbs some light,

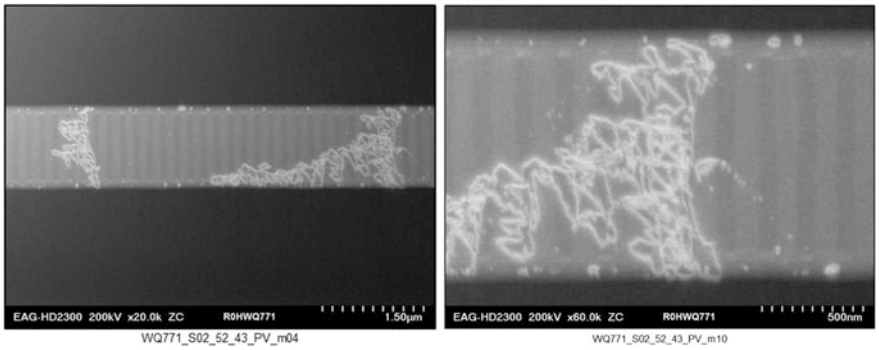


Fig. 10.19 Top view of the ridge shown in transmission electron microscopy, illustrating the growth and propagation of a dislocation network. These dislocation networks gradually (or suddenly) degrade the performance of semiconductor lasers

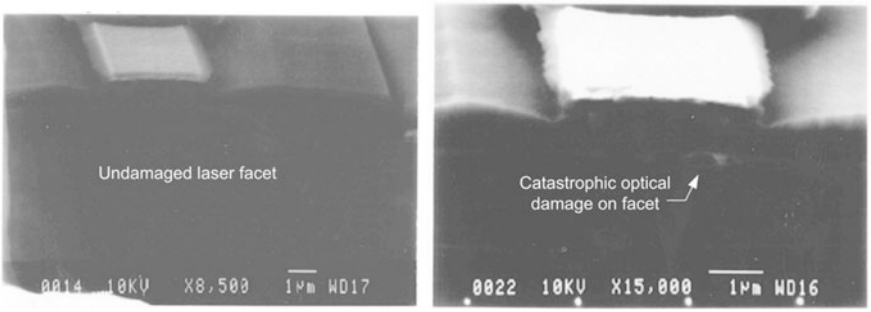


Fig. 10.20 Catastrophic optical damage on a laser facet

creating heat and causing defects on the facet, which leads to more absorption, and can lead to a positive feedback mechanism in which the facet rapidly melts, and the laser fails. A before and after picture of catastrophic facet degradation is shown in Fig. 10.20.

Sometimes these sudden failures are due to failures in the various external layers of oxide and metal that make up the device, or sometimes they just remain unexplained.

The third category is sometimes referred to as infant mortalities; occasionally, the devices fail suddenly after a few tens or less of hours of operation. Lasers are screened for this by operating devices at a highly stressed condition (high temperature or power) for a day or so, and then measuring the change in device active characteristics over the course of that time. Usually, these burn-in characteristics correlated to the long-term aging characteristics and can be used as a quick test of the device's expected reliability.

10.9.2 Definition of Failure

In this next couple of sections, it is the wearout failure mechanism that is being discussed. For analysis of reliability in a wearout mechanism, there has to be a definition of failure. Typically, the definition is based on an increase in operating current or decrease in power. For example, a 'failure' could be defined as 50% decrease in output power for a given current. Lasers all experience some level of degradation as they operate. The general operating requirements are not that the lasers maintain their initial specifications (for maximum threshold, minimum slope, and the like) over their lifetime; instead, the requirement is that they not degrade too rapidly.

10.9.3 Arrhenius Dependence of Aging Rates

From Fig. 10.18, the aging rate can be quantified in % per khr. This aging rate, AR, is a temperature-driven Arrhenius process, such as

$$\mathrm{AR} = A_0 \exp(-\Delta E_a / kT) \tag{10.5}$$

where k is Boltzmann's constant, T is the temperature (in K), and ΔE is the activation energy, which is typically of the order of electron volts (eV).

To find the particular activation energy, the aging rate can be measured at more than one temperature, and the relationship between the median aging rates at different temperatures can be used to determine the activation energy. Knowing the activation energy allows us to calculate the aging rate at a lower temperature from the measured aging rate at higher (accelerated) temperature.

Example: At 85 °C, the median aging rate of a set of samples is 1.2%/khr, and at 60 °C, it is 0.15%/khr. What is the activation energy?

Solution: $\mathrm{AR}_{85C}/\mathrm{AR}_{60C} = \exp((-\Delta E_a)(1/(8.6 \times 10{-}5\ \mathrm{eV/K})\ 1/(85 + 273) - 1/(60 + 273))) = 8$, so $\ln(8)(8.6 \times 10^{-5})$ * (1/(85 + 273) − 1/(60 + 273)) = 0.4 eV.

Values of 0.4–0.8 eV are typical of what is measured for wearout failures.

There are other 'acceleration factors' such as drive current and optical power which can also affect device degradation and wearout factors and are sometimes included in aging analysis. With models like this, they can predict expected aging rate at an operating point current from measured aging rates at one operating point (see Problem 10.4).

10.9.4 Analysis of Aging Rates, FITS, and MTBF

Analysis of aging rates starts by testing a set of samples at some accelerated condition, as shown in Fig. 10.17. The degradation of each device under test is measured, and a failure criterion is defined. From there on, the data set is analyzed statistically to determine the quantitative reliability of the device. Reliability is measured in Mean Time Before Failure (MTBF) and in Failures In Time (FIT), which is the total number of device failures in 10^9 device-hours of operation.

The statistical model which is usually used is that the MTBF and the aging rates are described by a lognormal process, in which the log of the relevant quantity follows a normal distribution.

The process is best illustrated with example.

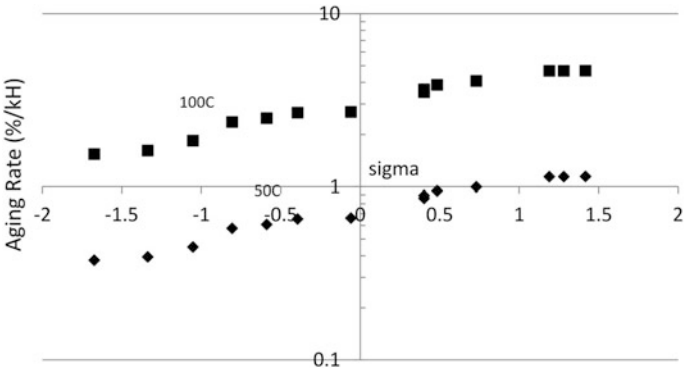


Fig. 10.21 Measured aging rates at 100 °C, along with calculated rates at 50 °C from activation energy and differences in temperature

To start with, let us look at the collection of aging rates of a sample of devices undergoing accelerated aging. Figure 10.21 shows the measured aging rates at 100 °C along with the rates calculated at 50 °C with Eq. 10.5. This plot is called a lognormal plot, in which the log of the aging rate (y-axis) is plotted against the standard deviation of the log(aging rate) function. On such a plot, the measured aging rates should be roughly linear and cross the 0 sigma at the mean.

Calculation of the reliability takes place at the hypothetical operating conditions, which in this case are uncooled devices hypothetically operating at 50 °C. The degradation rates are first calculated at the operating conditions though Eq. 10.5. Subsequent details of the analysis are illustrated in the example below.

Example: From the set of data shown in Table 10.4 (and graphed in Fig. 10.21), calculate the MTBF and the FITs. The devices are uncooled lasers with an expected lifetime of 10 years and an activation energy for aging of 0.4 eV.

Solution: Table 10.4 contains some calculated data and some measured data.

The left column is the measured degradation rate, which ranges from 0.5 to 2.7%/khr in this sample of 14 devices tested. The aging rate at 50 °C is calculated from Eq. 10.5 from the aging rate at 100 °C. The total aging (column 3) is the aging rate * khr in the specified ten-year lifetime. The power at fixed current is expected to decline by between 5 and 34% among this set of devices.

Degradation follows a lognormal distribution. The next step in the analysis is to take the natural log of the total aging (column 4) and measure its average and

Table 10.4 Aging data on some sample devices

Aging rate (100C) %/khr	Aging rate (50C) %/khr	Total aging (rate * khr)	Ln (aging)
0.5	0.07	6.36	1.85
0.6	0.09	7.63	2.03
0.7	0.10	8.90	2.19
0.8	0.12	10.17	2.32
0.9	0.13	11.44	2.44
1	0.15	12.71	2.54
1.2	0.17	15.25	2.72
1.54	0.22	19.58	2.97
1.54	0.22	19.58	2.97
1.61	0.23	20.47	3.02
1.84	0.27	23.39	3.15
2.36	0.34	30.00	3.40
2.48	0.36	31.52	3.45
2.67	0.39	33.94	3.52

standard deviation. In this case, the average is 2.75, with a standard deviation of 0.54.

The lognormal average is 2.75, which means a total average aging of exp(2.75) or 15.6%. The lognormal average aging rate is 15.6%/87.6 khr, or 0.17%/khr. If 'failure' is arbitrarily defined as a 50% decrease in output power, then MTBF = 50/0.17 = 294 khr, or about 34 years.

The ln of the failure condition (50%) is 3.9. In terms of standard deviation, that is about (3.9 − 2.54)/0.54 or 2.13 standard deviations from the mean.

The tabulated Gaussian Cumulative Distribution Function (CDF) is listed in terms of a dimensionless parameter Z, which is the number of standard deviations away from the mean. The cumulative number of failures (1 − CDF (2.13)) is 1.65%; 1.65% of the devices are expected to fail over their lifetime. Finally, the number of devices failing in a total of 10^9 device-hours can be determined by calculating how many devices are needed. The time of 10^9 device-hours represents 11,000 devices each operating for a lifetime (defined as 10 years, or 87.6 khrs). If 1.65% fail, that represents 188 individual failures in total 10^9 h, or 188 FITs.

As can be seen, both MTBF and FITs depend very strongly on both the median aging rate and on the distribution of aging rates. A narrow distribution (or low standard deviation) with a slightly higher average can give better reliability than a low average with a broader distribution.

Typical values range around 100 FITs (for uncooled devices) down to 10 or 20 FITs for cooled devices.

The process here takes months and months of test time. Usually, this detailed process is done once for a particular design, and then long-term aging results are done intermittently thereafter. Typically, reliability is monitored by short-term aging (a week or two) on a sample of devices from each wafer. Correlations have been established that allow degradation results over $\sim$200 h to project how the device will perform in long-term reliability.

Different variations on the methodology are followed by different companies. The reliability reports detailing the testing and analysis methodology, and the result in MTBF and FITs, are often used to convince the customer of the quality of the production process and the final product.

10.9.5 Electrostatic Discharge and Electrical Overstresses

In addition to meeting performance qualification, and reliability testing, lasers are tested to tolerate a certain amount of electrostatic discharge, which they could encounter during packing and handling. Typical test methodology involves testing a small sample of devices with progressive higher electrostatic discharge, until a device fails by demonstrating a significant shift in threshold, slope, or other operational characteristics.

The electrostatic voltage they can tolerate scales with the area of the device. For typical 200 μm edge-emitting devices, it is on the order of a thousand volts. Often, the devices are more tolerant of forward voltage than reverse voltage.

The failure of devices under electrostatic discharge can be dramatic. Figure 10.22 shows several different devices, after having exceeded the electrostatic discharge threshold. Even tested in reverse bias, the devices tend to be vulnerable at the facets, as seen.

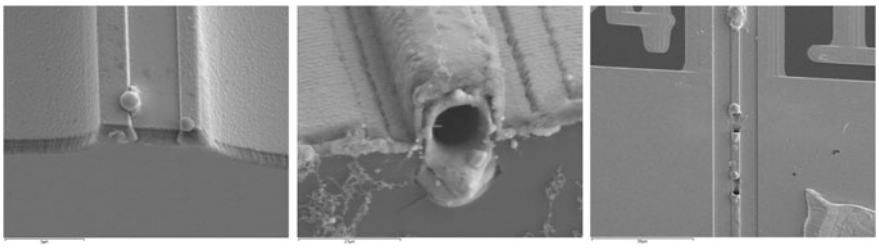


Fig. 10.22 Some examples of the damage caused by electrostatic discharge and electrical overstress, on the (left) mesa, (middle) facet, and (right) ridge of a laser

Typically, these are tested on a representative batch of devices to qualify a particular design. An electrostatic discharge threshold is set, saying that electrostatic voltages below this number do not harm the device.

10.9.6 Optical Overstress and Snap Test

To qualify the laser facet, devices are often put through an optical snap test, in which the current in the laser is ramped up briefly to a very high value. The purpose is to induce a very optical power level through the facet. Defects in the facet coating can lead to absorption in the light, creation of heat at the point and damage of the facet and facet coating.

10.10 Design for …

It is appropriate here to consolidate all the information discussed here and talk about how to design lasers. The focus of course is on the process, rather than the outcome, as requirements for lasers change continuously. By comparison of a couple of typical requirements, we can get a sense of what qualities can be important and the sorts of knobs that can be turned to determine the output.

The knobs that can be turned with laser design are subtle. Variables are things like device length and width, number of quantum wells, grating strength, doping profile. More detailed knobs include precise control of the grating profile. The outputs are things like power, performance over temperature, and bandwidth.

Table 10.5 Some tools useful for semiconductor laser modeling, as of 2019

Tool	Function
RSOFT	Various modules spanning component design (thermal, optical, electrical) through optical communication systems. Optical mode propagation and modeling
Crosslight Suite	
PhotonDesign Suite	
Lumerical Suite	
VPIphotonics	Modeling of laser output characteristics and transmission systems
TFCalc	Optical thin film reflectivity design tool
Laparex	Models output spectra of distributed feedback lasers
GLaparex	Models output spectra of distributed feedback lasers, including partial gratings, quarter-wave shifts

10.10.1 Design Tools

There are many design tools available which model laser performance to varying degrees of detail. Table 10.5 shows a couple that the author has used. In the author's experience, even the best tools do not perfectly capture a laser design; these are not digital silicon simulations. However, with experience they can indicate trends and be very useful in reducing the number of iterations necessary on a particular design.

While tools cannot replace actual experiment, neither can experiment replace tools. They are indispensable for narrowing down design space and low-cost iteration over many designs.

10.10.2 Design for High Speed Directly Modulated Lasers

In general, design for high speed looks something like the SEM shown in Fig. 9.6 and sketched in Fig. 10.23. Typically, the structure has many quantum wells (eight or more) to increase the differential gain. The separate confining heterostructure is relatively thin and often doped, to facilitate transport to the quantum wells. The cavity itself is short, to increase the photon density for given current injection while maintaining minimal capacitance and low photon lifetime.

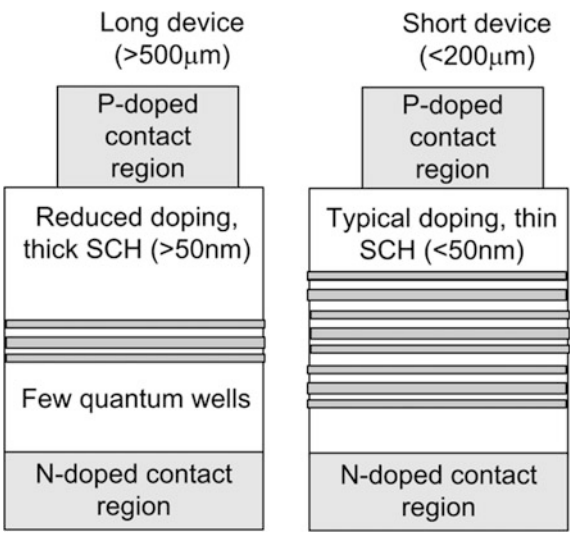


Fig. 10.23 Left, a qualitative sketch of a device design for high power with fewer quantum wells, larger separate confining heterostructure region, and reduced doping. Right, a qualitative sketch of a laser design for high speed, including many quantum wells for high differential gain, thin separate confined heterostructure regions for low transit time, and sometimes p-doped quantum wells

For buried heterostructure lasers, the blocking structure is designed as narrow as possible around the active region. It is often I-N, rather than P-N, to reduce parasitic capacitance.

All of these enable bandwidths well over 20 GHz in directly modulated lasers.

10.10.3 Design for High Power

High power lasers are important for silicon photonics and externally modulated devices. In silicon photonics, typically modulation and other functions are performed by the silicon, but the light has to be coupled onto and off of the silicon. Because the size of silicon waveguides is much different than laser waveguides, the loss involved in the coupling is quite significant, and so the laser power has to be quite high.

In high power, what matters is minimizing optical absorption and thermal effects. Absorption tends to scale with number of quantum wells, so typically higher power lasers are made with as few quantum wells as practical. The separate confinement heterostructure region is thicker than in a high speed device, as the larger separate confinement region keeps the optical mode away from the higher doped n- and p-contact regions.

Reduction in absorption has several beneficial effects. As a first order benefit, it allows more photons out per carrier in. As a second order benefit, since an absorbed photon often reappears as heat, reduced absorption means lower heating in the device. The heating and thermal rollover limit the ultimate power achievable out of the device.

Typically, the devices are much longer than high speed devices. A low absorption device is much easier to scale to longer lengths, and at longer lengths, the same power occurs at a lower current density and with less thermal rollover.

A sketch comparing typical high speed and high power lasers is shown in Fig. 10.23. The main point is simply to illustrate the idea that different applications lead to different basic design emphasizing different qualities. High speed lasers emphasize low capacitance and detuning for high differential gain; high power lasers emphasize low absorption and large size for low thermal rollover.

10.10.4 Design for Low Linewidth

A third major category of applications are the lasers used in coherent communication systems (which will be covered in Chap. 12). These lasers are generally used with external amplitude and phase modulation, and their major design consideration is very low linewidth (or, said alternatively, a very pure single mode spectral signal).

In general, these devices are external cavity or sometimes, integrated distributed Bragg reflector lasers. Integrated gratings tend to have much higher linewidths. Detailed design of these sorts of devices is not covered here, but good lasers for coherent communication tend to have linewidths well under a MHz, often under 100 kHz.

10.10.5 Design Over Temperature

Devices designed to be cooled, and operated at a fixed temperature, are typically designed for optimal performance at that operating temperature. Directly modulated distributed feedback lasers, which have higher differential gain when negatively detuned, are designed negative negatively detuned at the operating temperature. For devices designed to operate over temperature, the entire range of operation is considered. Typically, detuning at room temperature is set to maximize performance (e.g., threshold current) at the worst-case temperature, often high temperature.

Figure 10.3 indicates the general idea; as the temperature increases, for distributed feedback lasers, the lasing wavelength changes slowly ($\sim$0.1 nm/°C), while the gain peak beneath changes more rapidly ($\sim$1 nm/°C). The detuning between the gain peak changes with temperature, and this must be taken into account in designing a device to be used over temperature.

10.11 Summary and Learning Points

A. A major reason for distributed feedback devices is to obtain better quality long-distance transmission.
B. The quality of long-distance transmission is measured though a dispersion penalty, or difference in signal power required for same signal quality over fiber vs. back-to-back.
C. Typical specifications for dispersion penalty are 2 dB power penalty over operating conditions.
D. The temperature has a strong effect on the emission wavelength of Fabry-Perot semiconductor lasers. The bandgap and hence lasing wavelength increases by about 0.5 nm/°C.
E. The temperature also affects the emission wavelength of distributed feedback lasers, but only by about 0.1 nm/°C.
F. The temperature effect on emission wavelength can be used to tune the emission wavelength of the devices. For that reason, cooled wavelength division multiplexing devices usually can span two or three channels depending on the operating temperature selected.
G. Temperature also affects the DC properties, including the threshold current and the slope efficiency.
H. The effect of temperature on threshold is quantified by a phenomenological constant T_0 which quantifies the exponential dependence of threshold current on temperature.

I. For better high temperature performance, lower T_0 is better. Typical values for InGaAsP materials are about 40 K; typical values for InGaAlAs are 80 K or more. Because of that, InGaAlAs is the material of choice for uncooled devices.
J. Laser fabrication processes are outlined in Sect. 10.4.
K. Buried heterostructure devices and distributed feedback devices require regrowth (growth on patterned wafers) which makes them significantly more complicated than ridge waveguide devices, which do not require regrowth.
L. Buried heterostructures devices are generally slightly higher performance than ridge waveguide (in threshold and slope) but have additional parasitic capacitance.
M. Gratings in distributed feedback devices can be done with wafer-scale interference lithography, but are also done by wafer level e-beam lithography. E-beam lithography allows every device to have quarter-wave shifts and much improved single mode yield.
N. Vertical cavity devices are smaller, inherently single mode, and are easier to test on wafer; however, there is not yet a good commercial technology for longer wavelength (>900 nm) vertical cavity devices.
O. Device testing is done to guarantee that fabricated devices meet specifications. The testing is usually designed to find failing devices, or wafers, as early as possible.
P. In addition to tests of laser device characteristics, device reliability is also tested through accelerated aging, in which the laser is exposed to conditions far in excess of typical operating conditions in order to expose reliability failures early.
Q. Lasers have several failure modes, including infant mortality (sudden abrupt failures early), random failures (sudden failures which can occur at any time), and wearout failures which have to do with gradual performance degradation.
R. Laser aging rates follow a lognormal distribution, in which the log of the aging rates follows a normal (Gaussian) distribution.
S. Laser reliability is described by MTBF (Mean Time Before Failure) and FITs (Failures in Time, or failures in 10^9 device-hours).
T. Laser devices are also tested for electrostatic discharge tolerance, and for facet power handling capability.
U. Design considerations for typical laser devices used in communications include design for high speed, design for high power, and design for low linewidth. Each of these feature different trade-offs.

10.12 Questions

Q10.1. Dispersion is often compensated for in practice by dispersion-compensation links (lengths of fiber which are engineered to have a negative dispersion that will compensate for the positive dispersion experience on ordinary fiber.) Why can't these links be used to eliminate dispersion considerations altogether?

Q10.2. In fabrication described here, the grating used is buried within the device. Is it possible to put a grating on the surface of a device, and if so, what would be the advantages and disadvantages of it?
Q10.3. Would you expect a device designed with more highly strained layers to be more or less reliable than a device with less strained layers?
Q10.4. We note that the detuning reduces as the temperature reduces, to the point where a 20–30 nm detuning at room temperature can become 0 nm or negative at −20 °C. We also notice that the dynamics and high speed performance get worse as the detuning gets smaller. Do you expect this to be a problem in practice (e.g., for an uncooled device operating at an abandoned substation in the Arctic)?
Q10.5. What sort of problems would the reliability test not detect?
Q10.6. Why is the wearout failure rate in FITs so much less for dense wavelength division multiplexed devices than for uncooled devices?

10.13 Problems

P10.1. A typical specification for an uncooled telecommunication is $I_{th} < 50$ mA at 85 °C. If the T_0 of that particular laser is typically 45 K, what should the measured I_{th} be at 25 °C to be 50 mA or less at 85 °C?
P10.2. This problem discusses the maximum length that a 1480 nm laser with a chirp of 0.2 Å can transmit over optical fiber at 2.5 Gb/s, while maintaining a dispersion penalty less than 2 dB and optical loss of <30 dB. The fiber characteristics are losses of 0.5 dB/km and dispersion of 10 ps/nm/km at 1480 nm wavelength.

i. What is the maximum dispersion limited length?
ii. What is the maximum loss-limited length?
iii. 1.55 μm electro-absorption modulators typically can transmit up to 600 km dispersion limited transmission under the same conditions. What is their typical spectral width?
iv. How do 600-km transmitters overcome the fiber attenuation?
v. A far better natural choice for high speed transmission would be a directly modulated 1.3 μm device, with no dispersion. Why aren't 1.3-μm devices used for high speed long-distance transmission?

P10.3. Two different samples of ten devices each were put on accelerated aging tests, one at 85 °C and one at 60 °C. The one at 85 °C had a median aging rate of 2%; the one at 60 °C had a median aging rate of 0.4%. Calculate the activation energy appropriate for the accelerated aging.

P10.4. According to a JDSU White paper,[3] the random failure rate, F, is given by

$$F = F_0 \exp\left(-\frac{E_a}{k_n}\left(\frac{1}{T_j} - \frac{1}{T_{\text{op}}}\right)\right)\left(\frac{P}{P_{\text{op}}}\right)^n \left(\frac{I}{I_{\text{op}}}\right)^m, \tag{10.6}$$

where the subscript op is at the operating conditions testing, P is the optical output power, and I is the current. Take $m = n = 1.5$. If the FIT rate due to random failure at the tested condition of $T = 85°C$, $I = 50$ mA, $P = 2$ mW is 5000, calculate the FIT rate at $T = 60$ °C, $P = 2$ mW, $I = 35$ mA.

P10.5. A population of devices has an lognormal average rate of −2.9 (a rate of 0.055) and a lognormal standard deviation of 0.55 at its nominal operating temperature of 25 °C. Calculate the FITs in a 25-year lifetime and the MTBF.

P10.6. In the text, we state that the shift in lasing wavelength in distributed feedback lasers is 0.1 nm/°C. What fraction of that is due to thermal expansion of the lattice (for InP, the thermal expansion coefficient is 4.6×10^{-6}/°C)?

[3] http://www.jdsu.com/productliterature/cllfw03_wp_cl_ae_010506.pdf, current 9/2013.

Laser Communication Systems I: Amplitude Modulated Systems

11

The gull sees farthest who flies highest.
—Richard Bach, *Jonathan Livingston Seagull*

Abstract

Optical communications have gone to levels that were barely imagined twenty years ago. OC48 (2.5 Gb/s) is a thing of the past, and current state of the art commercial modules can deliver 100–400 Gb/s onto a fiber. In this and the following chapter, we outline how these rates have been achieved, both in high-speed device development and in an overview of the modern coherent communication formats. In this chapter, we give a brief history of optical communications, from where things were back in the mid-1990s to where they are today, and the evolution from directly modulated to externally modulated to coherent formats. We will also go into the details of the various amplitude-dependent formats and describe some of the physical mechanisms used for external modulation. We also briefly describe photodetectors used for communications.

11.1 Introduction

In this and the next chapter, we fly high and look at the communication system into which the laser goes. In the previous chapters, we have covered many of the details that define semiconductor lasers in communications: we have talked about stimulated emission, material families and their plusses and minuses, distributed feedback devices and many, many other topics. Beautiful and interesting these topics are, but in fact, they are specific details of the semiconductor lasers for communications and subordinate to the requirements of the communication systems.

© Springer Nature Switzerland AG 2020

D. J. Klotzkin, *Introduction to Semiconductor Lasers for Optical Communications*,
https://doi.org/10.1007/978-3-030-24501-6_11

We introduce the reader to the many different sorts of communication formats used in optical communication system. The requirements of the formats actually drive the individual laser requirements. Presenting it here is a good overview and review of the potential of semiconductor lasers in communications.

In addition to talking about the formats, we would also like to provide a working overview of the major technologies (in addition to lasers) that enable the specific formats. We are moving beyond laser modulation into the realm of communications. After being in the weeds for the previous ten chapters, let us soar above and examine the entire system.

The approach will be the following. First, we will talk about the various techniques for increasing communication bandwidth and the various technologies which enable them. The techniques are mostly forms of multiplexing to put a greater number of individual wavelength channels onto a fiber. Then, we will talk about various communication formats which let higher data rates transmit at lower noise on each given channel.

11.2 Evolution of Optical Speed

The diagram of Fig. 11.1 below indicates the evolution of optical speed over time with a picture of the enabling technologies.

The question to be asked and answered in this chapter is 'How did this happen'? How did we go from data rates of 2.5 and 10 Gb/s to fiber data rates at more than a Terabyte (Tb/s)?

11.3 Evolutionary Changes

Most of the answer is provided in the chart in Fig. 11.1, with this chapter filling in some details. They represent several paradigm shifts in the goal of achieving ever-greater speed.

In the 1990s and early 2000s, the focus was on development of single devices with modulation speeds of 2.5 and 10 Gb/s, which met the current need for fiber bandwidth. These directly modulated lasers provided sufficient bandwidth for the time, and there was always in the back of the mind of the industry the idea that many channels at these speeds could be multiplexed for an automatic bandwidth multiplier.

In the 2000s, this multiplexing explosion happened, and dramatic increases in fiber bandwidth were achieved by wavelength division multiplexing. In this way, fiber bandwidths of up to 1 Tb/s were accomplished.

However, individual directly modulated lasers ran into limits associated with dispersion as well as speed. Speed was intrinsically limited by the laser physics discussed in Chap. 8. Because direct modulation inherently introduced a chirp into

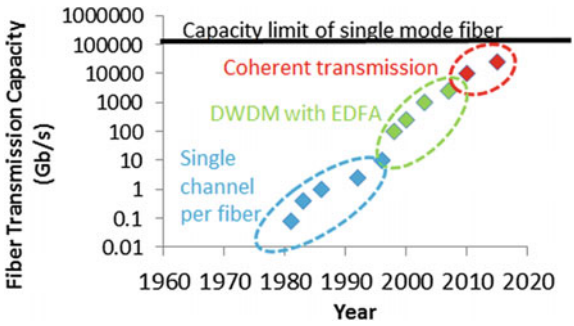


Fig. 11.1 Optical fiber data rate vs time along with the dates of the enabling technologies. The dramatic increases enabled both by dense wavelength division multiplexing, and by coherent detection, are clearly indicated. After *NTT Technical Review*, vol. 15, no. 2, Feb, 2017

the optical signal, the dispersion-limited transmission distance was also limited to roughly 100 km for 2.5 Gb/s.

In an effort to overcome both the intrinsic speed limit and the dispersion limit, externally modulated technologies were explored. These technologies included electroabsorption modulators and Mach–Zehnder modulators and offered higher modulation speeds with lower dispersion penalties. These technologies used a laser source that was continuous wave (CW) (unmodulated) coupled into a modulator.

After 2010, the focus shifted toward encoding. Borrowing techniques from coherent encoding from wireless technologies, information was encoded in both the amplitude and the phase of the optical signal. This allowed much more information to be communicated with the same physical symbol modulation speedrate. Methods based on these which require recovery of the optical phase are referred to as optical coherent communication methods.

The development of laser communications then can be loosely divided into several epochs:

- The directly modulated epoch, in which the laser itself was modulated; roughly the 1990s.
- The wavelength division multiplexing epoch, in which many more channels of directly modulated devices began to be put on to the fiber; roughly late 1990s-early 2000s. Development also included externally modulated devices, which took individual transmitters from 10 to 40 Gb/s with reasonable distance.
- The coherent epoch, in which both the phase and amplitude of the optical signal carry information, and many more bits are packed into each change of the physical characteristics.

These next chapters will give an overview of these developments.

The details of directly modulated devices have been covered in Chap. 8. Below we will cover multiplexing and then external modulation, and in Chap. 12, describe coherent networks.

11.4 Multiplexing

In previous chapters, in the discussion of laser modulation speed, we showed some bit patterns at 10 Gb/s generated by an individual device. Currently, transmission through a single-mode fiber can be of the order of Tb/s. In the section that follows, we discuss some of the simple methods in which that can be achieved.

11.4.1 Wavelength Division Multiplexing

The compelling advantage to optical communication is the enormous bandwidth that can be fit through a conveniently sized optical fiber. Because the wavelength is so small, the fiber can be small; again, because the wavelength is so small, the electromagnetic frequency is very high, and so an enormous amount of bandwidth and information can be transmitted on each wavelength.

An examination of the fiber characteristics in Fig. 1.3 shows the minimum loss at 1550 mm but relatively low loss from 1480 to 1620 nm. It also shows minimal dispersion at 1310 nm but relatively low dispersion around 1310 nm. If each wavelength can carry roughly the order of 10 Gb/s of information, the simplest way to increase bandwidth is to transmit information on many different wavelengths.

To enable different systems to work together, the International Telecommunication Union has established standards for sets of wavelengths used for communication. The standards are illustrated in Table 11.1. There are actually two different standards in use. The Coarse Wavelength Division Multiplexing Grid (CWDM) is intended for use with uncooled lasers. The spacing between channels is nominally 20 nm starting at 1271 nm. This spacing is large because (as we have seen) laser wavelength can drift significantly with temperature.

The denser grid of Dense Wavelength Division Multiplexing (DWDM) is intended for temperature-controlled lasers. The channel is specified by frequency, not wavelength: the allowed frequency values are given by

$$f_n = 193.1 + n * \Delta, \quad (\text{THz}) \tag{11.1}$$

where f_n is the frequency of a given channel, Δ is the frequency spacing, and n is an integer determining the channel number (positive, negative, or 0). For these tables, the relationship between frequency and wavelength is given by

Table 11.1 Wavelength bands for CWDM (a, specified by ITU 652.2), and DWDM (b, specified by ITU 652.1)

a. CWDM channel spacing[a]	
Wavelength (nm)	Frequency (THz)
1271	235.871
1291	232.217
…	…
1671	179.409

b. DWDM channel spacing (Δ *of 100, 25* GHz)[b]		
Wavelength, 100 GHz (nm)	Wavelength, 25 GHz (nm)	Frequency (THz)
…	…	…
1551.7208	1551.7208	193.200
	1551.9216	193.175
	1552.1225	193.150
	1552.3234	193.125
1552.5244	1552.5244	193.100
	1552.7254	193.075
	1552.9265	193.050
	1553.1276	193.025
1553.3288	1553.3288	193.000
…	…	…

[a]CWDM spacing is in wavelength, from 1271 to 1671 nm in 20 nm steps
[b]DWDM specification is in frequency, centered at 193.1 THz and width channels at 193.100 THz + n * Δ, where Δ is the assigned spacing, and n is a positive or negative integer. The channel spacing Δ can be as low as 12.5 GHz

$$f = c/\lambda = 2.99792458 \times 10^8 \,\mathrm{m/s}/\lambda \tag{11.2}$$

where c, of course, is speed of light in vacuum. Table 11.1, below, shows some of the values for the WDM grid.

A major advantage to semiconductor lasers is the ability to readily make lasers with so many different, precisely targeted wavelengths. By engineering the gain peak of the quantum wells, and the pitch of the grating, both gain and feedback at any of these wavelengths can be achieved. Hence, many of these wavelengths can be easily packed into a single fiber.

In addition to creating these many wavelength sources, the technology to combine signals of multiple wavelengths, and manipulate them individually, is necessary for wavelength division multiplexing systems. Below, we outline some technologies that are used to combine different wavelengths on and off of fibers.

11.4.2 Wavelength Division Multiplexing and Demultiplexing

Wavelength division multiplexing requires the ability to combine and separate multiple wavelengths onto and off of a fiber.

One marvelous device used for multiplexing and demultiplexing is an arrayed waveguide grating. A diagram of it is shown below. At the device level, a fiber with mixture of wavelengths will steer each wavelength onto a different fibers on the output end. For example, all the 1551.7 nm light in a fiber containing many wavelengths is coupled into one port, all the 1552.5 nm light is coupled into the next port, and so on. Operated in reverse, a particular wavelength can be coupled to the output fiber by putting that wavelength on its designated fiber on the right side. The picture below, Fig. 11.2, illustrates the concept.

The basic principal of operation is somewhat like a wavelength-dependent diffraction pattern. The input light from (1) is equally coupled onto the each of the waveguides in (3). However, the waveguides in (3) have slightly different path differences, so the array at the decoupling region in (4) the different wavelength have different phases. These paths serve to focus the coherent light at different colors into different locations.

It is similar, conceptually, to using the chromatic aberration of a lens in a positive way. Just like a camera lens with chromatic aberration focuses light of different colors to different spots, these arrayed waveguide does the same, only at the focal point of the array waveguide are waiting to collect the different colors. This can be used in either direction to multiplex onto a single fiber or demultiplex from an input fiber into multiple channels.

The drawback is coupling losses. Transitioning from a fiber to a waveguide to a fiber has associated losses that quickly add up. Light has to get from an optical waveguide on the order of 10 μm onto typically a semiconductor-based waveguide on the order of 1–2 μm, and then back again, and so losses can be on the order of 5 dB.

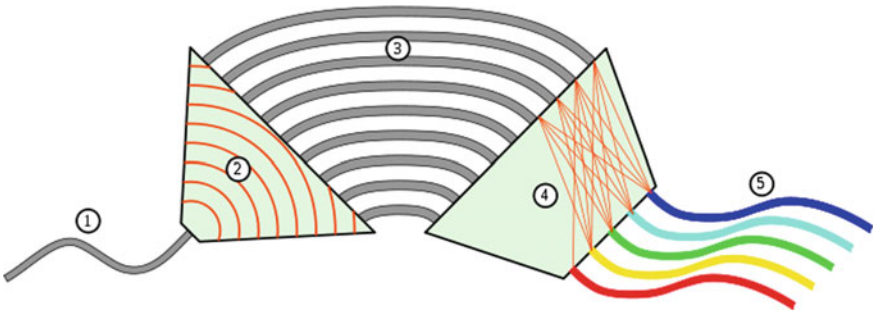


Fig. 11.2 Arrayed waveguide. Incoming light from 1 is split into a number of individual lasers, and then free-space-coupled to an array of fibers with specific distance and spacing such that the different colors of light couple into different fiber ports. Image from Wikipedia, https://en.wikipedia.org/wiki/Arrayed_waveguide_grating, current, 1/2019

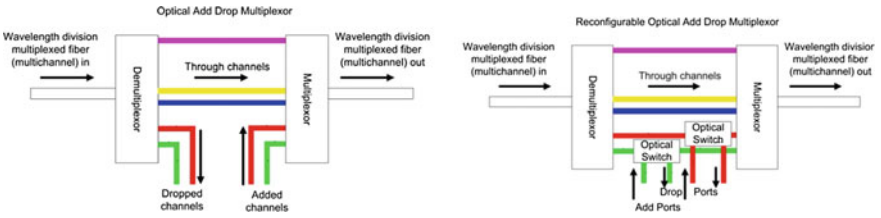


Fig. 11.3 Left, optical add drop multiplexor, and right, reconfigurable add drop multiplexor. The devices allow switching of wavelengths conveniently onto or off of optical fiber

11.4.3 Optical Add Drop Multiplexors

In addition to splitting light into multiple ports by wavelength, the second thing that is needed is the ability to switch individual wavelengths on to or off of the primary fiber. An illustration of the concept is shown in Fig. 11.3. An optical fiber carrying many wavelengths is not usually conveying them all between two points, but instead drops off a wavelength at one location, picks up another at different location, and so on. Imagine it as an interstate highway with many entrance and exit ramps.

This routing of wavelengths on and off of a fiber is accomplished by a device called an Optical Add Drop Multiplexer (OADM). The individual wavelengths can be switched on or off as needed. For example, in the left picture, two wavelengths are permanently dropped at the location (maybe a college campus) and outgoing data from the campus then switched back onto the fiber using the same two colors. This is an example of static OADM.

The picture on the right shows a dynamically reconfigurable optical add drop multiplexer. These allow each node on the network to be changed as traffic conditions change. They require controllable optical switching, where wavelengths can be routed on or off depending based on input control signals. These switches can be based on microelectromechanical mirrors, tunable filters, or other switch mechanisms.

11.5 Overview of Amplitude-Modulated Communication

After all the multiplexing we can reasonably do, we are left with a bottleneck. The physics of lasers limits direct modulation to something no more than 20–30 Gb/s, and so other techniques have to be explored in order to the keep the bandwidth increasing.

The simplest communication format is the directly modulated digital on–off keying that we encountered in previous chapters. By modulating the current into the laser, we can change the light level. One light level can be interpreted as a ‘1’, and

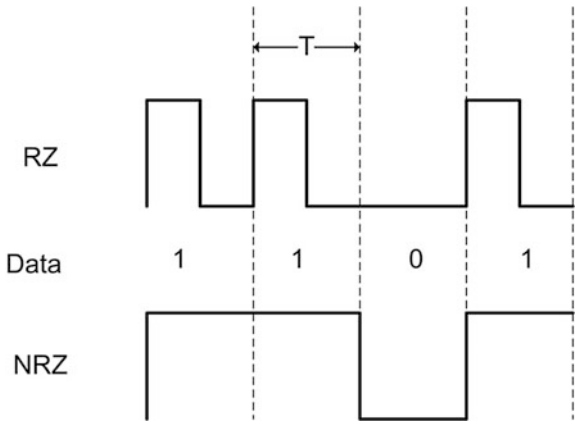


Fig. 11.4 NRZ versus RZ format, RZ format returns to zero after each bit, while NRZ encodes each symbol in a single bit period

one as a '0', and information can be transmitted. However, that runs flat into the ultimate, intrinsic limit on laser communication (the laser modulation speed).

There are techniques that have been evolved to pack more information onto an optical signa, and those will be covered below, after definition of the appropriate vocabulary.

11.5.1 Definitions for Amplitude Modulation Formats

Simple amplitude modulation formats are classified as NRZ (non-return to zero) and RZ (return-to-zero). The diagram of Fig. 11.4 illustrates them. A RZ format returns to zero after each bit (whether it is a one or a zero), while in an NRZ format, each pulse is a '1' or a '0'. In RZ format, the optical signal switches during a '1' but remain constant for a '0'.

The tradeoffs are clear; the RZ format generally encodes the clock rate and is easy to distinguish, but at the cost of half the symbol bandwidth. In general, in optical communication systems, bandwidth is too valuable and NRZ is used.

11.5.2 Bits Versus Symbols

The distinction between bits and symbols is very important as we go on to discuss communication formats. A bit is a unit of information, and bit rate is expressed in units of bits/s (or usually in an optical context, gigabits/second, Gb/s). A symbol is something which changes to communicate information, and a symbol rate is measured in baud.

$$\text{bits/s} = (\text{symbol/s}) * (\text{bit/symbol}) \tag{11.3}$$

If every '1' or '0' changing represents a single bit, then bit rate and baud rate are the same. However, if the information is encoded so that every physical change represents more (or less) than one bit, then they are not. Information transfer is measured in bits/sec; symbol change is measured in baud. A simple way to visualize baud is in the potential number of symbol transitions per second.

In Fig. 11.4, the NRZ format has one bit for every transition, and so the bit rate equals the symbol rate. The question of bit rate and symbol rate of RZ format is covered in the problems, but below is a related example.

Example: in a NRZ modulation format, the level varies between high and low, so the bit rate of information, in bits/s (usually Gb/s, in an optical context) is equal to the symbol rate, 1 Gbaud.
One convenient format for digital information is called Manchester encoding, in which a '1' is represented by a 1–0 transition, while a 0 is represented by a 0–1 transition.[1] The advantage to such encoding is that there is a transition at least once every bit, so the information content of the signal is centered at the data rate. For the Manchester encoded signal shown in Fig. P11.5, if each clock period is 1 μs (clock rate 1 MHz), what is the symbol rate, in baud and the information transfer rate, in bits/s?

Solution: As can be seen, one bit of information is sent every clock period. Therefore, the bit rate is 1 Mb/s. Each bit can have two transitions (if it is 1 1 1); therefore, the symbol rate is (up to) 2 M transitions-per second or 2 Mbaud. Manchester encoding is one of the few formats in which the symbol rate is more than the bit rate.

Since there are two transitions/bit,

1 Mb/s = 2 Mbaud * 1 bit/2 transitions

The advantage to Manchester encoding is that the signal encodes its own clock very cleanly and has its spectrum centered around the baud rate, not a DC signal. These characteristics make the signal easy to receive and filter, which is some return for giving up half the channel bandwidth. One disadvantage to Manchester encoding is that because the information content is centered at the bit rate, not up to the bit rate, it generally requires a higher bandwidth receiver.

[1] as you can see in the figure, there are two conventions: the Thomas convention has a '1' represented by a '1'-to-'0' while the other convention has a '1' represented by a '0' to '1' transition.

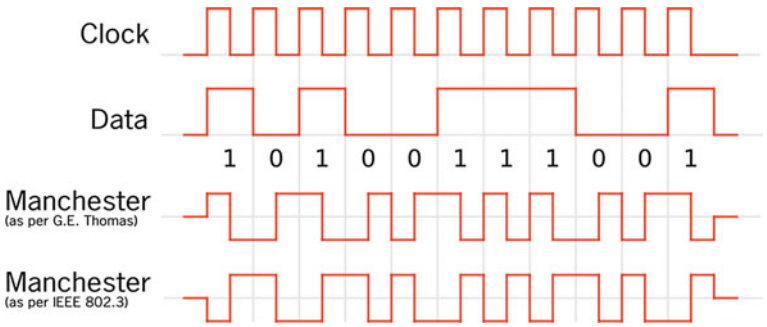


Fig. P11.5 Demonstration of data which is Manchester encoded. From Wikipedia, https://en.wikipedia.org/wiki/Manchester_code, current 1/2019

In on–off keying, or the direct modulation concept discussed in Chap. 8, there is 1 bit/symbol, and so the baud rate and the symbol rate are the same.

11.5.3 Pulse Amplitude Modulation

There is one small adaption that can be made to squeeze more bandwidth out of direct modulation. Pulse amplitude modulation (PAM), as illustrated in Fig. 11.6, extends digital on–off keying into more than just two levels (on and off). If the receiver can detect more than one level reliably, then each level can encode more than one bit.

In general, in a system with n levels, the number of bits is given by

$$\text{bits} = \ln_2 \text{levels} \tag{11.3}$$

The figure illustrates how bits are encoded in a four-level system. Of course, the tradeoff is in receiver signal-to-noise ratio. In a sense, it is a step backwards into the analog world, when a level could encode high-precision information, and not just a '1' or a '0'. However, practically speaking, many receivers can resolve four levels under reasonable conditions, and so it is an effective way to increase communications bandwidth.

As of 2018, this PAM-4 format is widely used in high-speed data centers, often in packages with multiple lasers. It can realize 50 Gb/s/channel with lasers with bandwidths of roughly 20 GHz.

However, it does impose signal-to-noise limits, as it requires being able to distinguish between four levels, instead of two. Conceptually, it is easy to imagine more than four different levels, with a consequently higher number of bits/symbol.

This is largely the limit of what is currently used for directly modulated amplitude modulation.

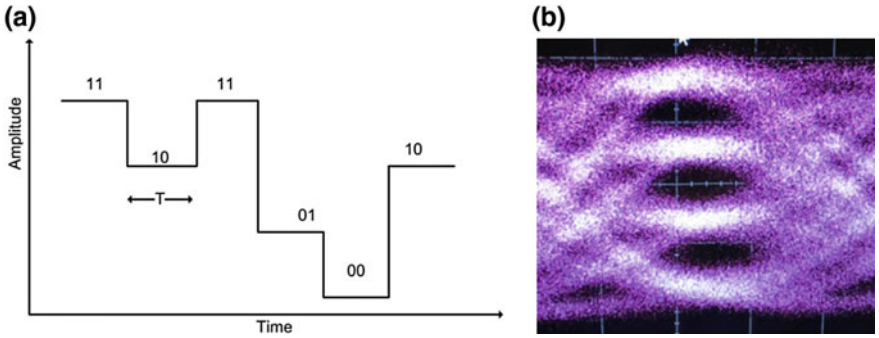


Fig. 11.6 **a** Illustration of a PAM-4 time-domain signal, in which each of the 4 levels encodes two bits. **b** A real 50 Gbaud PAM-4 eye pattern signal, showing a single laser transmitting at 100 Gb/s

11.6 External Modulation

External modulation is the next step for higher speed and longer distance transmission. With external modulation, rather than modulating the laser directly, the continuous wave (CW) laser light is put through a second device which modulates it through a variety of mechanisms. This offers two advantages.

First, in terms of speed: direct modulation of lasers is inherently limited by capacitance and intrinsic effects to numbers typically not much greater than 20GHz[2]. This limit is inherent to direct laser modulation; other modulation technologies have different, often higher, inherent speed limits. By using an external modulation technology, that limit can be increased.

The second reason is dispersion. The dispersion penalty, which fundamentally limits transmission for amplitude-modulated systems, depends on the chirp (which is the change of wavelength as the device is modulated) through Eq. 10.1. Direct amplitude modulation of the laser inherently changes the chirp, but external modulation does not. Therefore, dispersion limitations for external modulation systems are less.

There are some disadvantages to external modulation. For example, the need to couple light into a separate device leads to an insertion penalty. This insertion penalty is a built-in loss of power going from one waveguide (the laser) to another (the modulator) and can be several dB. The second disadvantage is the increased complexity and cost of the external modulator.

Let us now discuss some of commonly used external modulation mechanisms.

[2]The number, 20 GHz, can be quibbled with. There have been reports of devices made with different techniques achieving over 50 GHz bandwidth, but they have not yet proven practical direct modulation devices.

11.6.1 Quantum-Confined Stark Effect

Before discussing this external modulator based on the quantum-confined Stark effect as a device, let us describe the modulation of absorption through bias through quantum-confined Stark effect mechanism. Semiconductors generally absorb light depending on their bandgap and the wavelength of the light. What is needed is an electrically controllable mechanism that can be rapidly modulated to change the absorption through change in an electrical signal, and the quantum-confined Stark effect does exactly that.

The effect is illustrated in Fig. 11.7, along with a qualitative illustration of the change in the absorption with bias. Imagine a quantum well, like the ones discussed in Fig. 4.4. For an unbiased structure, the absorption starts to occur between the first two confined subbands, at an energy level of $E_{\text{bandgap–unbiased}}$. For an electrically biased structure, the details of the bandstructure are altered, and a tilt, due to the applied bias, is applied to the quantum well potentials. The electron subband sinks into the deeper side of the quantum well, while the hole subband floats up into the higher half of the quantum well. When the subbands are calculated in this new structure, they are closer together, and absorption starts to occur at the new lower energy, $E_{\text{bandgap\_bias}}$.

If the specific wavelength of light is right around the intersubband spacing of the quantum wells, the absorption can be modulated readily by changing the bias. This is the principle behind the electroabsorption modulator.

The application of this effect to a device is shown in the next section.

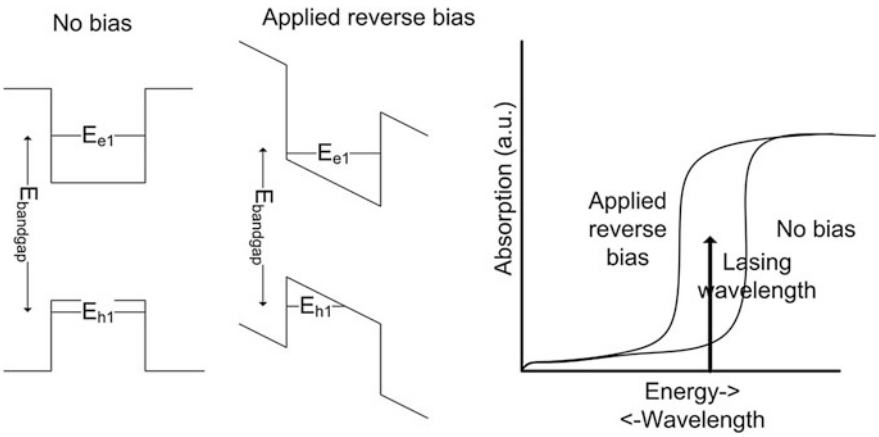


Fig. 11.7 Illustration of the quantum-confined Stark effect. A quantum well with no bias across it has its bandgap, and absorption, start fundamentally at the intersubband spacing as shown. With an applied bias, the electron subband sinks a little lower and the hole subband goes up a little higher, and the bandgap between them decreases. The graph on the right shows absorption versus energy or wavelength for two different applied biases. If the wavelength to be modulated is placed right around the bandgap, small changes in bias can have large changes in absorption and amplitude of the output light

11.6.2 Absorption Modulation Through the Quantum-Confined Stark Effect

Figure 11.8 illustrates the basic structure of an integrated laser modulator based on the quantum-confined Stark effect. On the left is a standard distributed feedback quantum well laser, and on the right is a quantum-confined Stark effect modulator. Both the laser and modulator are grown on the same wafer and often use the exact same quantum wells (which is possible if the grating pitch is used to correctly choose the laser wavelength). An electrical isolation (achieved by etching or implantation) isolates the two distinct sections of the device.

To transmit data, a signal is modulated onto the modulator portion of the device.

An example of some typical data for a structure like Fig. 11.8 is shown in Fig. 11.9.

The two figures illustrate some aspects and tradeoffs of these integrated laser modulators. Generally, the lasing wavelength is positively detuned from the gain peak, so zero-bias on the modulator is transmissive through the modulator region. These devices typically must be temperature controlled, as the relationship between the lasing wavelength and the absorption peak is quite critical. However, as can be seen in Fig. 11.9, with a proper configuration, an extinction ratio of 10 dB or more can be achieved with a few negative volts reverse bias. The modulation bandwidth is typically limited only by the capacitance of the modulator.

One tradeoff, however, is the optical loss of the modulator. The absorption does not abruptly go from zero to some large number. Depending on the wavelength chosen for the transmitter, there is some absorption in the transmissive-state (on the order of a dB) with a corresponding power-level reduction.

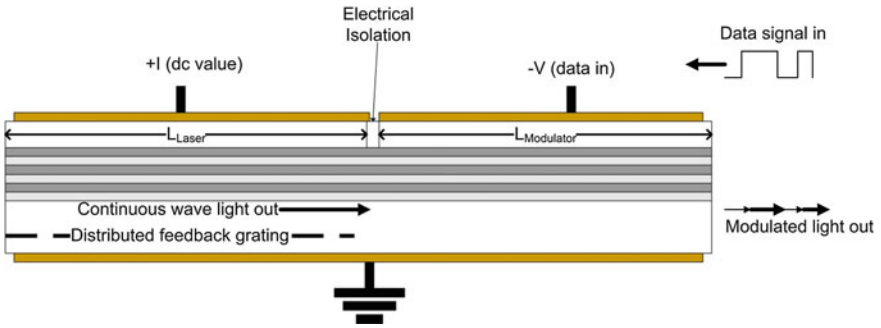


Fig. 11.8 Integrated laser modulator. The left side of the structure provides a continuous level output light, while the right side modulates the light through the quantum-confined Stark effect. The two sides must be electrically isolated (as shown). The absorption modulation does not introduce chirp into the signal and is primarily limited in speed by R–C considerations and capability of the driver, but the overall structure is more complicated that the directly modulated laser and is difficult to realize uncooled

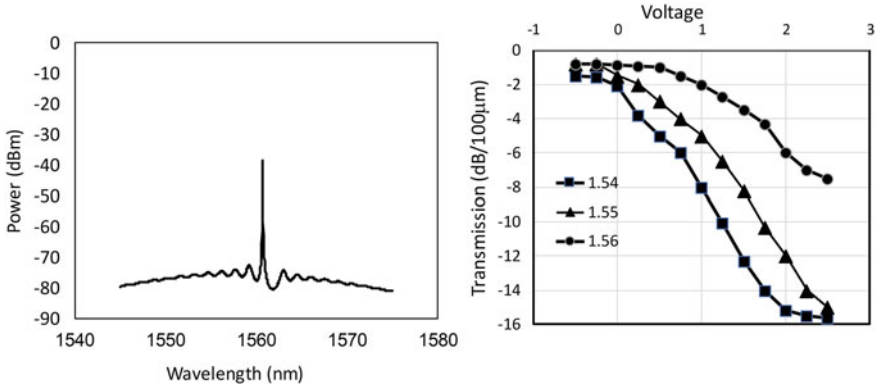


Fig. 11.9 Some data from an integrated laser modulator similar to the one pictured in Fig. 11.8 (after *IEEE J. Sel. Top. Quantum Electron.*, vol. 2, pg. 326). As the exact same quantum wells are used for both modulation and light generation, the light has been positively detuned (longer wavelength than) the gain peak. As shown in the figure on the right, application of reverse bias lowers the transmission of light through the structure

Example: In the modulator whose data is shown in Fig. 11.9, calculate the absorption (in/cm) for a wavelength of 1.55 μm at 2 V reverse bias. Note that the graph gives transmission for 100 μm modulator distance.

Answer: 10 log $(P_{out}/P_{in}) = -12$, so $P_{out} = 10^{-1.2} = 0.063$. The number $0.063 = \exp(-\alpha 0.01) \rightarrow \alpha = 276/\text{cm}$.

What is the insertion loss for a lasing wavelength of 1.54 μm? 1.55 μm? What is the extinction ratio for the two wavelengths at 2 V?

Answer: The insertion loss, defined as loss at 0 V bias, is roughly 1.7 and 0.7 dB at 1.54 and 1.55 μm, respectively. At 1.54 and 1.55 μm, the extinction ratio is loss (0 V)-loss(2 V), or 13 dB and 12 dB, respectively.

This integrated laser modulator has a place in amplitude-modulated transmission at moderate distances and high speeds. The need for speed at very long distances is met largely by coherent technologies that can be scaled to very long distances and moderate to high single-channel data speeds, as will be discussed in the next chapter.

11.6.3 Mach–Zehnder Modulator from Electooptic Materials

Another external technology which is used for optical modulation is the Mach–Zehnder interferometer, used as a modulator. Figure 11.10 shows a general Mach–Zehnder interferometer, in which the incoming optical signal is split between two arms and then recombined at the output. The optical path length between the two paths can be changed through application of an electrical signal. Hence when the optical signals are recombined, amplitude modulation is realized though either constructive or destructive interference of the two signals.

Realization of a Mach–Zehnder modulator requires a material and mechanism that can controllably shift the phase of the light by changing its refractive index. One mechanism is the electrooptic effect in which application of an electric field leads to changes in refractive index.

A canonical example of a material which has this effect is lithium niobate ($LiNbO_3$). When electrical fields are applied across a waveguide formed of lithium niobate, its refractive index undergoes a relatively large change. This change changes the phase of light propagating in the waveguide relative to an unbiased waveguide. Because the change in lithium niobate is relatively large, and it is otherwise a tractable material to fabricate, it is frequently used in optoelectronics. To realize full modulation and completely turn off the optical signal, the nominal amplitude has to be the same, while the difference in phase between the two paths of light has to be π radians.

Figure 11.11 shows the use of lithium niobate as it would be integrated into a device.

The relationship between index change Δn and electric field E in lithium niobate is given by

$$\Delta n = -\frac{1}{3} n_0^3 (rE) \tag{11.4}$$

where Δn is the refractive index change, n_O is the refractive index, and r is the electrooptic coefficient (usually in units of 10^{-12} m/V, or pm/V).

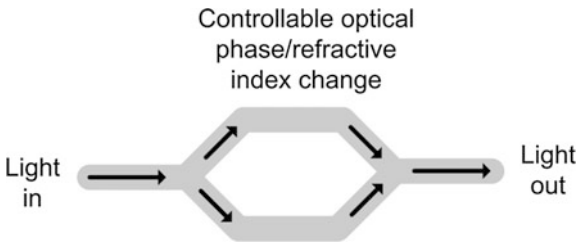
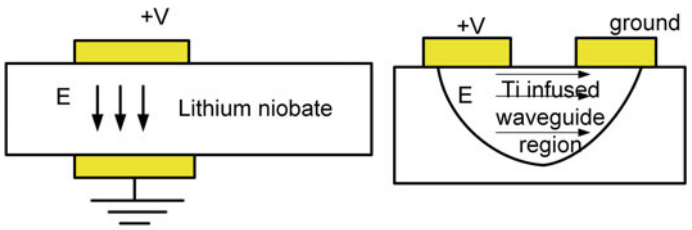


Fig. 11.10 Simple picture of a Mach-Zehnder interferometer. The light is split and recombined. The difference in phase between the two paths can realize constructive or destructive interference at the output, and create an amplitude-modulated signal. Some physical mechanisms allow control of the refractive index and hence the optical path lengths of the different paths to realized constructive or destructive interference, and hence amplitude modulation, at the output



$$\Delta n = -\frac{1}{2} n_0^3 (rE)$$

Fig. 11.11 Litihum niobate electrooptic material and waveguides. The useful property of lithium niobate is illustrated in the inset equation, where the refractive index changes with applied voltage. The righthand figure illustrates a basic lithium-niobate optical device. A waveguide is created by infusing Ti into the $LiNbO_2$, which increases the refractive index.

A typical refractive index of lithium niobate at 1.55 μm is 2.11, though this varies with Ti doping and with crystal orientation. Figure 11.11 shows the waveguides and electrodes of a typical lithium niobate device. Details of materials, fabrication, and devices vary, but generally, the lithium niobate material is formed into a waveguide and then sandwiched by electrodes. Application of a voltage shifts the refractive index of the material, and thus the phase as the light propagates a fixed distance.

A figure of merit for a fabricated modulator is V_π, the voltage needed to shift the phase by 180° and turn the transmitted signal from on to off. The following example illustrates this in more detail.

Example: Figure P11.12 shows two arms of a lithium niobate structure, both sandwiched by metal electrodes, as is typically used for optical communications. For the geometry and electrooptic phase coefficient shown in the figure, what is the V_π (the applied voltage needed to shift the phase of the light by π (180) relative to an unbiased waveguide).

Answer: First, let's calculate the total phase along an unbiased waveguide, the propagation constant k is 2π $(2.12)/1.55 \times 10^{-6}$, or 8.5894×10^6/m. Thus, the total phase change along a 500 μ the total phase change^6 (500×10^{-6}), or 4294.1 radians. A π phase change implies a total shift of 4291.6 radians. Since the length is equal, that implies k is $4291.6/500 \times 10^{-6}$, or 8.5832×10^6/m. Then $n = 8.5832 \times 10^6$ $(1.55 \times 10^{-6})/(2\pi) = 2.118$. The refractive index change needed is 2.120–2.118, or 2×10^{-3}.

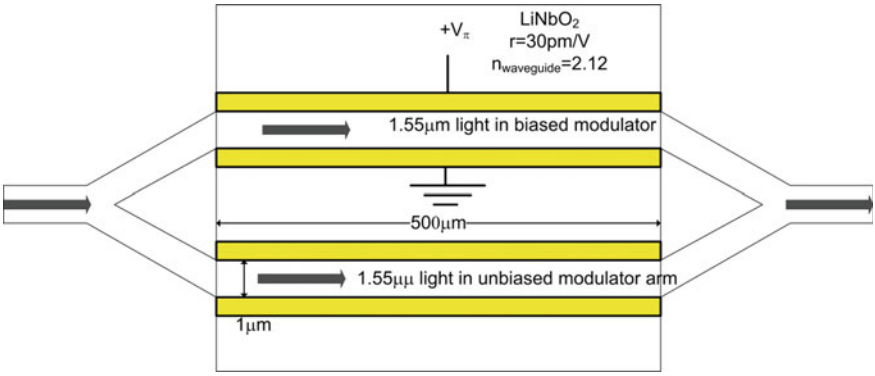


Fig. P11.12 Lithium niobate Mach–Zehnder modulator

Putting that into Eq. 11.4, $2 \times 10^{-3} = 1/3\ (2.12)^3 (E)(30 \times 10^{-12})$, or $E = 2.1 \times 10^7$ V/m. Given the 1 μm separation, $V_p = (2.1 \times 10^7)(1 \times 10^{-6}) = 21$ V. In order to shift the phase of the light by π radians, a voltage of 21 V has to be applied.

The value 21 V is a large number. As of this writing, commercial lab-grade lithium niobate phase modulators are available with V_π of <5 V. Both the geometry of the waveguide (particularly length) and the material property of the electrooptic coefficient figure into this V_π number.

These lithium niobate devices are relatively large compared to Si electronics or typical communication laser lengths. They are often on the millimeter scale, compared to semiconductor lasers (on the 100 μm scale) and electronics (on the submicron scale).

The splitting, recombining, and phase changes serve to realize optical modulation. One implantation of it is shown in Fig. 11.13, with different directions of electric field applied across each arm.

The figure shows a classical, Mach–Zehnder interferometer structure with a differential biasing scheme (such that one arm has a +V bias while the other arm has a −V bias). The light is split 50–50 at the input arm and a phase shift imposed on each arm. A differential configuration means the output is in phase with the input, but at a different amplitude. The details are explored in the following example.

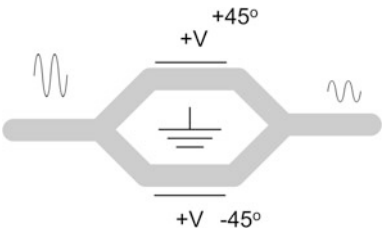


Fig. 11.13 Mach–Zehnder interferometer. The configuration of biases means that one arm is biased positively and one arm is biased negatively. Because of this, the output signal is in phase with the original signal but reduced in amplitude

Example: In the Mach–Zehnder shown in Fig. 11.13, one arm shifts the phase of the signal by 45° with respect to an unbiased arm, and the other arm, biased oppositely, shifts the phase by −45°. What is the amplitude and phase of the output, assuming a perfect 50:50 split at the input and the output?

Solution: Basically, it is ½ (cos (ωt + 45) + cos (ωt − 45)), which is simplified to 0.707 cos wt. The net effect of this Mach–Zehnder is to reduce the amplitude to 70% of the original amplitude, while keeping it in phase with the original signal. One of the properties of differential biasing is that it stays in phase with the original signal, which is important when used with a quadrature amplitude modulation scheme.

Here, the Mach–Zehnder interferometer is an amplitude modulator, and can be an alternative to electroabsorption modulation or direct modulation. Though it does not have the same fundamental speed limitation that directly modulated laser does, it suffers from its own problems. For example, a high extinction ratio (defined as before, as ratio between the maximum and minimum amplitude at the out) requires an exact split between the two arms. If the two arms split, for example, 55:45, the modulation signals will not cancel even if they are 180° out of phase.

11.6.4 Phase Shifting with Plasma Effect

There are other mechanisms which can shift the refractive index of a material based on some sort of biasing signal. One of them is based on increasing or decreasing the mobile carrier concentration in silicon (or other semiconductor) materials. Because

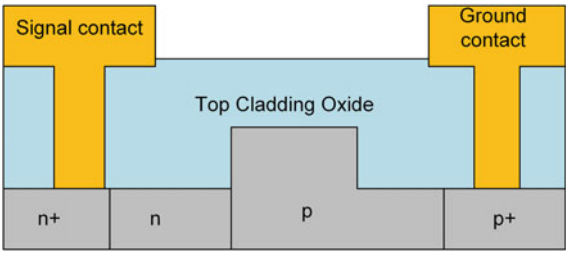


Fig. 11.14 Basic structure of a silicon optical modulator (after *IEEE Photon. Technol. Lett.*, vol. 24, pg. 234)

mobile carriers affect the refractive index, modulating the number of carriers (though carrier injection or depletion) changes the local refractive index. This is called the plasma effect, and it is strong enough to engineer Mach–Zehnder interferometers similar to those based on the electrooptic effect.

This effect allows the direct engineering of modulator and phase shifters on a silicon waveguide. A typical Si-base phase shifting structure is shown in Fig. 11.14.

When a bias is applied to the signal, a depletion layer appears between the n and p segments. The width grows or shrinks based on the bias. The optical mode sees the average refractive index of the waveguide. Its effective index and the phase of a wave traveling in that waveguide is changed as well. From a device perspective, it is similar to an electrooptic material, but the mechanism is different; electrooptic materials respond to electric fields, while the plasma effect depends on modulating carrier density.

Empirically, the change in refractive index depends on the change in carrier densities as

$$\Delta n = -(8.8 \times 10^{-22} \Delta N_e + 8.5 \times 10^{-18}\, \Delta N_h^{0.8}) \tag{11.5}$$

at a wavelength of 1.55 μm in silicon, where Δn is the change in refractive index, and ΔN_e and ΔN_h are the changes in carrier densities, in/cm^3. This will be explored further in the problems.

The ability to realize controllable phase shifters with Si allows devices like Mach–Zehnder modulators, add drop ports, multiplexers, and splitters to all be combined onto one chip, for enormous optical integration. The combination of lasers with silicon may actually be another transformative step in optical communications, enabling another order of magnitude increase in fiber bandwidth at even lower cost.

Regardless of how they are realized, Mach–Zehnder interferometers are important alternatives to direct modulation. They are crucial to coherent communication which will be discussed in Chap. 12.

11.7 Laser Linewidth

We have been talking about externally modulating continuous wave (DC) laser light. One clear advantage, and a major reason external modulation is preferred, is reduced laser chirp. However, even when driven with unmodulated current, the linewidth (or range of wavelengths) in a laser signal is non-zero. Let us return to lasers for a section to talk about laser linewidth.

Laser linewidth is important to directly modulated dispersion limited transmission, as we discuss here, and utterly crucial to coherent transmission, as we discuss in the next chapter.

First, for dispersion-limited transmission: one of the primary differences between directly modulated devices, as discussed in Chap. 8, and externally modulated devices, as discussed here, is the degree of chirp. Directly modulated devices inherently chirp (or change wavelength with modulation current) which limits their dispersion-limited transmission. These externally modulated devices do not have that chirp and so in general have much great dispersion-limited transmission distances. However, all lasers and semiconductor lasers, in particular, do have a finite linewidth (they are not perfectly monochromatic), and so there is still a limit for the externally modulated dispersion transmission in those devices.

Here, we give a brief overview of mechanisms for linewidth in semiconductor laser.

11.7.1 Inherent Laser Linewidth

For the coherent lasers will discuss in this and the next chapter, the most important criteria are not speed or power, but linewidth. A narrow linewidth means the laser precisely remains at its lasing wavelength or frequency, and more importantly in the context of coherent communication, that its phase does not jump around.

Intuitively, it seems like light emission from stimulated emission should be extraordinarily narrow, as every stimulated emission photon remains in phase with other photons. Schalow and Townes, however, long ago predicted that even processes driven perfectly by stimulated emission would not have a zero linewidth.[3] There are random processes which are ongoing which add noise to the output signal. In particular for semiconductor lasers, spontaneous emission, which occurs randomly, couples into the lasing mode. Transmission or absorption of photons is also on a photon-by-photon basis and subject to statistics, so it, in addition, gives rise to noise. These factors give rise to a 'minimum' semiconductor laser linewidth $\Delta\nu_{\text{min}}$,

$$\Delta\nu_{\text{min}} = \frac{v_g h\nu R_{sp}\alpha_m}{8\pi P_0} \tag{11.6}$$

in which $h\nu$ is the energy/photon, v_g is the group velocity, α_m is the modal loss in per length units, and P_0 is the total power in the optical model. The item with most

[3] A. L. Schalow, C. H. Townes, "Infrared and Optical Masers", Phys. Rev. 112 (6), 1940 (1958).

uncertainty is R_{sp}, the total amount of spontaneous emission (in/s) which couples into the lasing mode.

This lower limit value in semiconductor lasers is typically on the order of 10 s of kHz. It is much higher than theoretical limit set by Schalow and Townes, and also higher than seen in other sorts of laser.

Example: Estimate the fundamental laser linewidth in a device with an optical loss of 10 cm^{-1}, at a lasing power of 10 mW and wavelength of 1.3 μm.

Solution: The word 'estimate' gives wide latitude in the solution process: let us begin. Typical group velocities v_g are about 1×10^{10} cm/s. The energy/photon $h\nu$ is about 1.24/1.3 = 1 eV/photon = 1.6×10^{-19} J/photon. The only quantity remaining is the amount of spontaneous emission coupling into the lasing mode, R_{sp}.

Total spontaneous emission is n_{th}/τ (from the rate equations); picking typical numbers of $10^{18}/cm^3$ as threshold current density 10^{-11} cm^3 as active region volume, and 10^{-9} s as carrier lifetime gives $R_{sp-total} = 10^{18} * 10^{-11}/10^{-9}$ or 10^{16} recombinations/s total. The typical assumption is 10^{-4} or so of spontaneous emission photons couple into the lasing mode, so $R_{sp} = 10^{12}$/s.
Multiplying it all together gives, $10^{10} * 1.6 \times 10^{-19} * 10^{12} * 10/(0.01 * 8 * 3.14) = 600$ kHz. It is very rough number, but it gives a sense of the order of magnitude of the linewidth.

Here, the point is to appreciate that the random factors of spontaneous emission and optical loss add to phase variation and set a lower limit for the linewidth.

11.7.2 Linewidth Enhancement Factor

In addition to linewidth variation driven largely by spontaneous emission, there is an additional factor with semiconductor lasers. In semiconductor lasers, the gain and the refractive index are directly coupled together through the plasma effect. Fluctuations in carrier density are clearly fluctuations in optical gain. Less intuitive, though, is that fluctuations in carrier density are also fluctuations in refractive index. The optical mode sees a fairly large volume around the active region, and the index of each material is modified by the density of mobile charge. Hence, fluctuations in mobile charge affect both gain and refractive index, and shifts in refractive index lead directly to shifts in wavelength.

The effect of this is that amplitude, and phase or frequency noise are coupled together, quantified through a parameter called the linewidth enhancement factor or Henry's α factor. Semiconductor lasers are noisier than other mediums because of the strong coupling between amplitude and phase.

The mathematical definition for the α factor given below as is

$$\alpha = \frac{4\pi}{\lambda}\frac{\partial n}{\partial g} = -\frac{4\pi}{\lambda}\frac{\partial n/\partial N}{\partial g/\partial N} = -\frac{4\pi}{\lambda}\frac{\partial n/\partial I}{\partial g/\partial I} \tag{11.7}$$

where $\partial n/\partial N$ and $\partial g/\partial N$ are refractive index and gain variation induced by carrier density injection; respectively, and $\partial n/\partial I$ and $\partial g/\partial I$ are variation by current.

The phenomena are illustrated in Fig. 11.15 which shows a common method of measuring the linewidth enhancement factor. Two measurements of the lasing spectra below threshold are taken at slightly different currents. The gain increases, shown by the increase of the peak amplitude, and the index changes, shown by the shift in peak location at the different current. From these two spectra, numbers for the linewidth enhancement factor can be calculated.

Although this measurement was taken below threshold on a Fabry-Perot laser, the effects of the linewidth enhancement factor are seen above threshold in both Fabry-Perot and distributed feedback devices. Devices with higher linewidth enhancement factors have greater dispersion penalties under direct modulation. In addition, even under direct current, this linewidth enhancement factor fundamentally impacts the laser linewidth.

Including the linewidth enhancement factor, the linewidth of a laser is given by

$$\Delta v = \Delta v_{\min}(1 + \alpha^2) \tag{11.8}$$

Example: This example is based on the data of Fig. 11.15. If the optical gain extracted (via the Hakki-Paoli method of Chapter 7) at 5.2 mA is 12/cm at a wavelength of 1304.00 nm, and at 5.4 mA is 16/cm at a wavelength of 1303.95 from a cavity with a group index of roughly 3.5 at 5.2 mA, find the linewidth enhancement factor.

Solution: The actual change in current is irrelevant; we can just calculate change in gain and change in refractive index. The change in gain is 4/cm. The nominal wavelength is 1304 nm. Since the basic wavelength-index relationship is $L = m\lambda/n$ with m an integer and L is the cavity length. Hence, $\lambda_{5.2}/\lambda_{5.4} = n_{5.4}/n_{5.2} = 1303.95/1304 = 0.9999961$; $\Delta n = 1.4 \times 10^{-4}$. So the linewidth enhancement factor is

$$\alpha = -\frac{4\pi}{1.303 \times 10^{-4}}\frac{1.4 \times 10^{-4}}{4} = 4.4$$

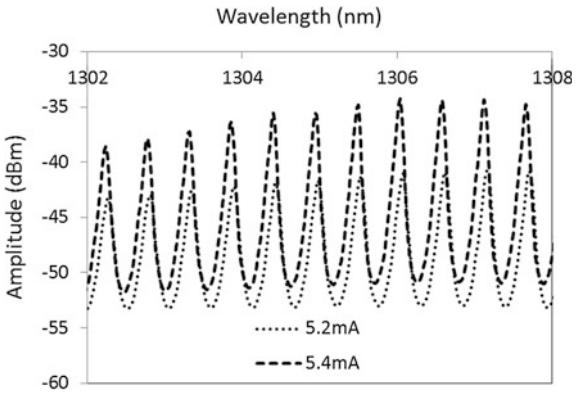


Fig. 11.15 Laser spectra taken at two different currents slightly below threshold, with shift in peak wavelength indicating a change in refractive index and shift in amplitude peak-to-valley indicating a change in gain

Typical values for α are 2–5 in multiquantum well devices. This factor depends on the details of the laser, including its quantum confinement structure, the details of the heterostructre design and the location of the wavelength with respective to the gain peak. In particular, devices which are negatively detuned tend to have smaller α factors.

The mechanisms and rough numbers for linewidth enhancement are illustrated in Fig. 11.16.

11.8 Direct Detection Receivers

Though this text is focused on lasers and the transmitter side of optoelectronics, it is worthwhile here to spend a page (or a page and a half) briefly describing the optoelectronic devices which receive optical signals and convert them back into electrical impulses. In the next chapter, we will be implementing systems which can detect the phase of an optical wave; it is useful to understand how the basic, amplitude photoreceivers work, and their limitations. Here, will talk about the most common p-i-n photodiodes used as detectors.

The basic photoreceiver component is a p-i-n photodiode, as pictured in Fig. 11.17. In a common arrangement, the photodetector is top-illuminated, in a center area ringed with a p-contact ring. The light is absorbed in the intrinsic region and generates an electron hole pair. The built-in field, or an applied reverse bias, sweeps the carriers toward the n- and p-contacts, where they are collected.

These p-i-n diodes have responsivities of the order of one electron hole pair per photon.

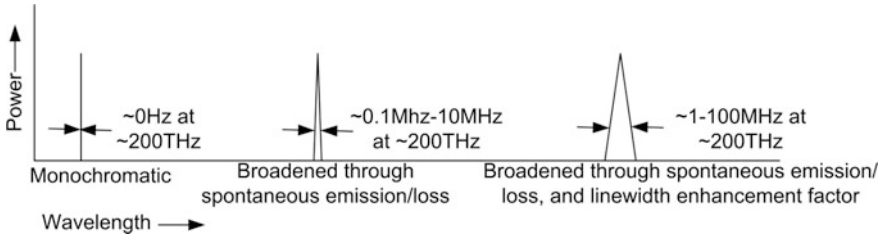


Fig. 11.16 Mechanisms and rough amounts by which the linewidth in semiconductor lasers is broadened

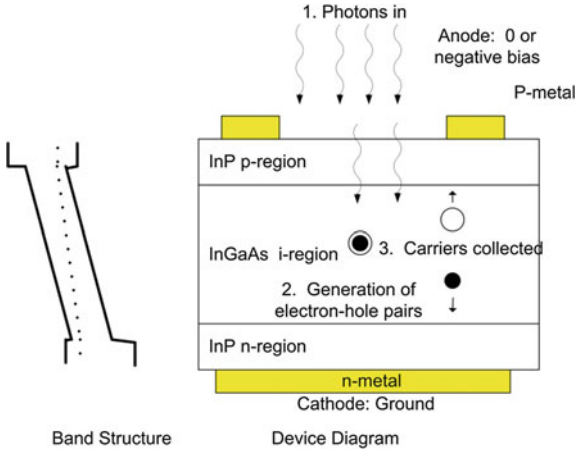


Fig. 11.17 Left, the band structure of a typical p-i-n photodiode Middle, device diagram of a photodetector made of InP contacts sandwiching an InGaAs absorption layer

There are two fundamental speed limitations associated with them. The first has to do with the thickness of the absorption layer, and transport of carriers to the contact, where they are collected. Once the electron hole pair is generated in the middle of the undoped i-region, it needs to be collected by drifting to the appropriate contact. This drift time is set by the drift velocity (which depends on applied reverse voltage and carrier mobility), and the distance, set by the thickness of the absorption region. The transit time limit is expressed as

$$f_{-3\,\mathrm{dB}-\mathrm{transit}} = \frac{1}{2\pi \frac{d}{v}} \tag{11.9}$$

where d is the distance traveled by electron or hole to be collected, and v is the velocity. The velocity depends on the electric field through the mobility, but it is usually considered as the saturation velocity of the carrier in that material. In this

simple expression, we have lumped the velocity of the electrons and holes into a single typical velocity v.

The second has to with the capacitance of the diode. The large the diode area, and the thinner the active region, the higher the capacitance and the longer the RC time constant. Quantitatively, the RC limit can be expressed as

$$f_{-3\,\mathrm{dB}-RC} = \frac{1}{2\pi RC} \tag{11.10}$$

in which R and C are the series resistance associated with the photodiode and photodiode circuit. The capacitance, in particular, includes the parallel plate capacitance of the diode structure. Some aspects of the frequency limits are explored in the problems.

These kinds of diodes can be fabricated with bandwidths approaching 100 GHz. They can see the amplitude of the envelope of the optical signal, but not the peaks and valleys of the 100 THz optical field. To determine the phase of an optical wave requires interference techniques, as we explore in the next chapter on coherent communication.

11.9 Summary and Learning Points

In this chapter, we have stepped a few paces back from how lasers work and put them into the context of how optical transmission systems work. Here, we have covered amplitude modulation formats and mechanisms, and in the next chapter, we will talk about coherent communication.

A. Some important definitions: a 'symbol' is the thing which changes in data transmission (such as an optical amplitude), while a bit is a unit of information.
B. Directly modulated lasers are limited in speed to something about 50 Gbaud/s (currently) due to fundamental laser limitations, but despite that, fiber bandwidths greater than Tb/s have been realized.
C. The increase in link transmission bandwidth has come about through a variety of techniques, including multiplexing, more bits/symbol and higher speed external modulation, and (to be covered next chapter) coherent transmission techniques
D. Wavelength division multiplexing puts many channels at different wavelengths on the same fiber and multiplies the link bandwidth accordingly.
E. Even though each individual channel is limited in speed, the transmission windows around 1550 and 1310 nm each have many channels and wavelengths which can transmit independently on the fiber.

F. These different wavelengths have to be multiplexed onto and off of a single fiber. This is typically done with arrayed waveguide gratings as part of an optical add drop multiplexor.
G. One extension of digital modulation which can extend the bandwidth is pulse amplitude modulation. In this format, the optical signal comes in several levels, and each level represents more than one bit of information.
H. In the simple, on–off, non-return-to-zero amplitude modulation, the symbol rate is the bit rate; in other forms of transmission, such as pulse amplitude modulation, the bit rate can be different than the symbol rate.
I. An alternative to direct laser modulation is external modulation, in which continuous wave laser light is coupled into a different device which does the modulation.
J. One external modulator is based on modulated absorption in quantum wells through the quantum-confined Stark effect. This allows integrated lasers and modulators (fabricated on the same wafer, often with the same quantum wells) to be fabricated.
K. Another external modulator is based on a Mach–Zehnder interferometer. These modulate by splitting the signal light into two segments, changing the phase in one path, and then recombining the two light segments. By modulating the phase, constructive or destructive interference creates modulated amplitude signals.
L. Mach–Zehnder modulators can be realized in with electric fields through the electrooptic effect. These are done external to the laser often with lithium niobate.
M. Mach–Zehnder modulators can also be realized through charge injection or removal through the plasma effect. These modulators can be integrated onto semiconductors, such as InP or Si.
N. Phase shifting as introduced here is crucial for coherent communications.
O. Lasers have an inherently non-zero linewidth due to random fluctuations from spontaneous emissions and loss terms. Semiconductor lasers also have strong coupling between amplitude and phase variations through the linewidth enhancement factor, which limits their inherent linewidth to numbers of the order of 10 s of kHz.
P. P-i-n photoreceivers are the workhorses of optical receivers. Typically, they can have bandwidth up to 100 GHz, sufficient to see the envelope of the optical wave but not to observe the phase directly.

11.10 Questions

Q11.1 What new technologies enabled the increase in bandwidth from the 1990s?
Q11.2 What is the difference between bits and symbols?

Q11.3 For amplitude shifted formats, there are several different ways of putting bits into symbols. Explain the differences between on–off keying (OOK) and pulse amplitude modulation (PAM), return-to-zero (RZ) and non-return-to-zero (NRZ), formats. What is Manchester encoding?
Q11.4 List two alternatives to a directly modulated laser for amplitude-modulated optical transmission.
Q11.5 Build a table comparing directly modulated, quantum-confined Stark effect modulated, and Mach–Zehnder modulation in terms of speed, complexity, insertion loss, dispersion, and cost.
Q11.6 What factors limit the linewidth of semiconductor lasers?
Q11.7 What factors limit the speed of a p-i-n diodes?

11.11 Problems

Q11.1 A pulse amplitude modulated with four levels has a symbol rate of 25 Gb/s.

(a) What is the bit rate?
(b) Sketch the eye pattern seen by an oscilloscope looking at this wave form. Indicate the period.

Q11.2 A wavelength division multiplexing system is being designed for a temperature range of 0 to 50 °C. Roughly how many channels can be fit into the standard O-band from 1290 to 1330 nm? (assume the typical thermal shift for a distributed feedback laser of 0.1 nm/°C)
Q11.3 In the RZ transmission system shown in Fig. 11.4, if the clock period is 100 ns,

(a) What is the bit rate?
(b) What is the symbol rate (in Gbaud)?

Q11.4 For the electroabsorption modulator material shown in Fig. 11.9,

(a) Find the modal absorption at 1.56 μm at a bias voltage of −2 V.
(b) If the material is an 8 quantum well active region with a confinement factor of 10% to the quantum wells, find the material absorption at 1.56 μm at a bias voltage of −2 V.
(c) How long would the device have to be to have an extinction ratio of 10 dB? What would the insertion loss be at that length?

Q11.5 Find an expression relating the change in index, the modulator length, and the wavelength, for a π phase shift in one arm of a Mach–Zehnder modulator.

Q11.6 Design a Mach–Zehnder interferometer, using the material in the example of Sect. 11.5.3, that would have a V_π of 1 V.

Q11.7 The figure below shows a schematic of a silicon-based Mach–Zehnder interferometer.

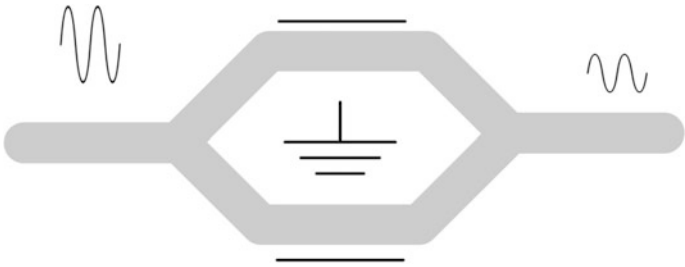


Fig. P11.18 Sketch of a Si-based differential Mach-Zehnder modulator

Each arm of the modulator is based on a p–n junction in which current can be injected.

a. If a change in charge density of 5×10^{17}/cm for both electrons and holes is realized, evaluate the change in index at 1.55 μm.
b. Find the amplitude of the output signal with the charge density given with a length of 200 μm.
c. What length should the modulator be in order to introduce a phase shift of π between the two arms?

Q11.8 Another mechanism to introduce phase change is through change in temperature of the material. In silicon, the index change with temperature is given by

$$\frac{dn}{dT} = 1.86 \times 10^{-4}/\mathrm{K}$$

where n is index, and T is temperature in Kelvin.

a. What temperature change is necessary to get an index change of 10^{-3}?
b. With that change, what is the length necessary to realize a Mach–Zehnder interferometer with a potentially infinite extinction ratio?
c. Why do you think this mechanism is not used to realize modulation for optical transmission?

Q11.9 A Mach–Zehnder interferometer such as that shown in Fig. 11.16 has a 55–45 split at the input arm. What is the maximum achievable extinction ratio if the two arms are 180° out of phase?

Q11.10 (a) In the laser example of 11.7.1, what would the actual linewidth of the laser be if the device had a linewidth enhancement factor of 4?

(b) Looking at Eq. 11.6, suggest a couple of ways the laser could be engineered or operated for a lower linewidth enhancement factor.

Q11.11 This problem explores the frequency limits of the p–i–n detector. In the figure shown, the area of the diode is 25 μm^2 and the thickness of the i-region is 1 μm.

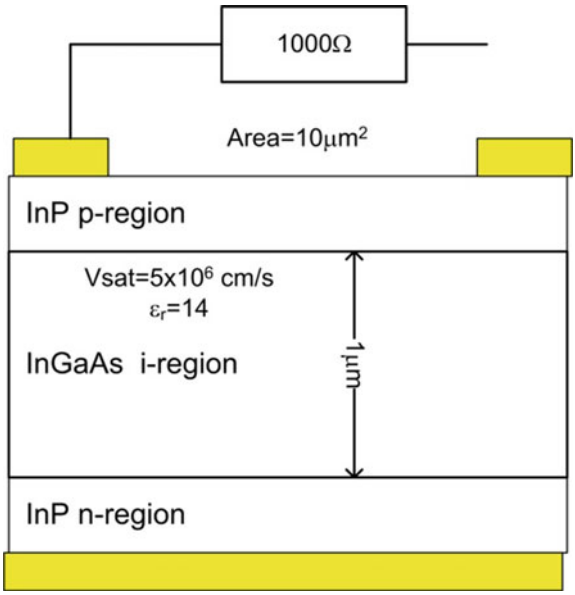


Fig. P11.19 Sketch of a typical p-i-n photodector showing width of the intrinsic region, geometry, and resistance of the contacts

(a) For the parameters shown, calculate the RC limit and the transport limited bandwidth of the photodiode. Assume the capacitance is just given by the photodiode capacitance, the carrier absorbed in the center of the band, and a single 'velocity' as a model for both hole and electron velocity.
(b) Suggest one way to improve the RC limit without affecting the transport limit.
(c) Assuming that it is limited by the lower of the two limits, find the optimum thickness d to achieve the highest bandwidth.
(d) Apart from increasing the capacitance, what disadvantage is there in reducing the thickness of the absorber region, in terms of responsivity?

Coherent Communication Systems 12

...and a great and strong wind rent the mountains...and after the wind an earthquake... and after the earthquake a fire... and after the fire, a still small voice...
—Kings 19:11–12, King James Version

Abstract

In this chapter, we outline the modern coherent communication formats and how they are realized. These coherent communication formats enable long-distance and high-speed transmission. We also discuss the sources of noise in optical communication systems and the relationship between signal to noise, bit error rate, and communication format.

12.1 Introduction

In the previous chapter, we described as far as has been currently realized with amplitude modulation and some technologies for amplitude modulation apart from direct laser modulation. However, there is an entirely different method of information transmission with light, focused on what is called coherent detection. In this method, both the amplitude and the phase of the (200+ THz!) light signal with respect to some reference communicate information. By measurement of the phase through combining the signal with a reference, a 'still, small voice', hidden in the phase of the optical signal can emerge. Using all the information contained in the photons, rather than just signal amplitude, dramatically improves the amount of information that can be communicated optically. But this relies on seeing light differently than we have so far.

© Springer Nature Switzerland AG 2020

D. J. Klotzkin, *Introduction to Semiconductor Lasers for Optical Communications*,
https://doi.org/10.1007/978-3-030-24501-6_12

12.2 Phasor Representation of Light

12.2.1 Reminder: Phasor Representation of Electrical Signals

If you are now using this book as a textbook, there have been, hopefully, many connections to your prior classes such as semiconductors for p–n and Schottky junctions and ohmic contacts, and physics for waveguide modes, for example. This chapter is a solid connection to circuits (which the author also has the privilege of teaching).

In the discussion of sinusoidal (AC) circuits, the conventional technique is to describe AC signals as phasors. For example, a voltage with a particular frequency and amplitude, and a phase of some angle θ with respect to some reference signal, is described as $A\angle\theta$, as shown in Eq. 12.1. As an example, the current through a 1 H inductor due to a 1 V signal at 120 Hz (377 radians/s) and its phasor representation are both given in Fig. 12.1 on the phasor diagram.

$$f(t) = A\cos(\omega t + \theta) \Rightarrow A\angle\theta = Ae^{i\theta} = A\cos\theta + iA\sin(\theta) \tag{12.1}$$

at a frequency ω. Typically, the vector $\mathbf{A}$ is plotted on the complex plane.

Both the amplitude and the phase of the signal matter. For example, because the current and voltage in an inductor are out of phase by 90°, inductors consume no power.

12.2.2 Phasor Representation of Optical Signals

In the case of circuits at moderate frequency, observing the time-domain current or voltage signal is generally possible with an oscilloscope, and the phase of the signal with respect to some arbitrary reference can be easily determined. For optical signals, techniques to generate and detect the phase are more indirect, but the concept is the same.

One more bit of terminology before we start talking about optical transmission: a phasor signal has a component $I(t)$ in phase with the reference and a quadrature component $Q(t)$ 90° out of phase with the reference. Hence, an optical signal $L(t)$

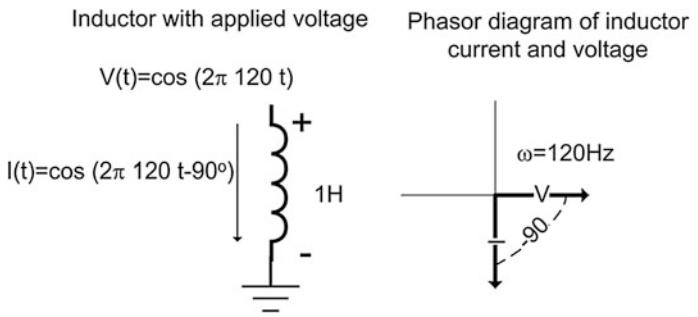


Fig. 12.1 An illustration of a phasor diagram for an *I–V* curve. Each vector represents both the amplitude and the phase, relative to the reference voltage

$$\begin{aligned} L(t) &= A\cos(2\pi 192\,\text{THz}\,t + \theta) \\ &= A\cos(2\pi 192\,\text{THz}\,t)\cos(\theta) + A\sin(2\pi 192\,\text{THz}\,t)\sin(\theta) \end{aligned} \tag{12.2}$$

in which A is the optical amplitude has as its phasor representation,

$$L = I + iQ \tag{12.3}$$

(with the frequency understood to be 192 THz), with the values

$$\begin{aligned} I &= L\cos\theta \\ Q &= L\sin\theta \end{aligned} \tag{12.4}$$

Transmission formats that rely on the phase of the light are called coherent communication formats, and they are remarkably powerful. In general, I and Q are functions of time ($I(t)$ and $Q(t)$) and change at the symbol rate of GHz scale, much slower than the optical signal at hundreds of THz.

Figure P12.2 and the associated example illustrate decomposing a time-domain signal into its I and Q components.

Example: Fig. P12.2 shows three different signals at optical frequencies plotted in the time domain. Estimate the time-domain expression for the signal as a single cosine term, and illustrate both the signal and the two references in the phasor domain. Find the I and Q magnitudes.

Solution: From the figure, we need to estimate the relative amplitude and phase of the signal. The amplitude looks

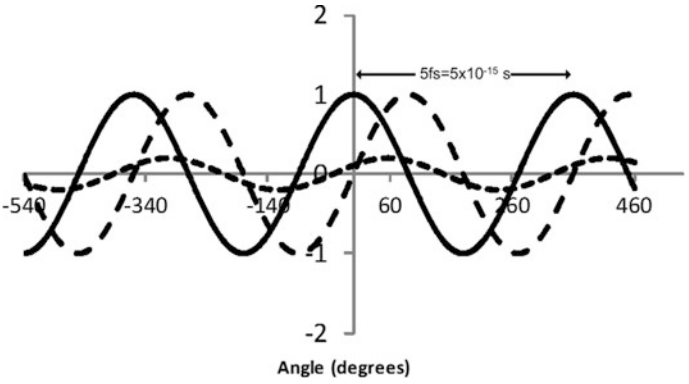


Fig. P12.2 A sinusoidal signal in the time domain, along with the cosine and sine references. The full amplitude sin and cosine are given as 'reference signals,' and the signal is an arbitrary optical signal. In coherent communication, generally both the amplitude and phase of the incoming signal are used to convey information

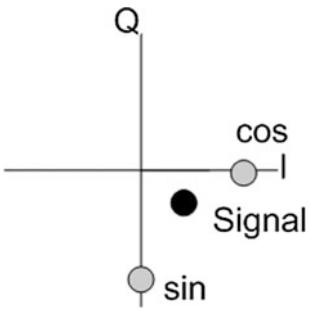


Fig. P12.3 A phasor diagram of the signal. The I and Q magnitude would be $I = 0.25 \cos 45$ and $Q = -0.25 \cos 45$

like it is about a quarter of the reference signals, so call it 0.25; the peak comes between the cos and sine signal, so the absolute difference in phase is 45°. The angle is negative (leading signals come first in time and have their maximums earlier than the signals they lead. Therefore, the signal leads the sine wave and lags the cos wave). On a phasor diagram, it would look like Fig. P12.3.

12.3 Phasor Descriptions of Coherent Optical Transmission

With the idea of phasor representation of optical communication, let us show some of the formats used for coherent communication.

12.3.1 Binary (and More) Phase Shift Keying

One of the simplest coherent techniques is called binary phase shift keying (BPSK), along with its logical extensions into higher numbers of phases. Instead of changing the amplitude of the optical signal, the phase is changed while the amplitude is fixed. This can be done among as many different phases as we can confidently detect. In Fig. 12.4, it is illustrated with both two and four different phases.

In all of these systems, the receiver has to be sensitive enough to detect and discriminate between the different symbols. Both amplitude and phase noise may be present, and the system has to be able to resolve the symbol in the presence of noise.

12.3.2 Differential Phase Shift Keying

One variation that is appropriate to mention is differential phase shift keying. With differential phase shift keying, instead of a particular phase conveying a bit, a

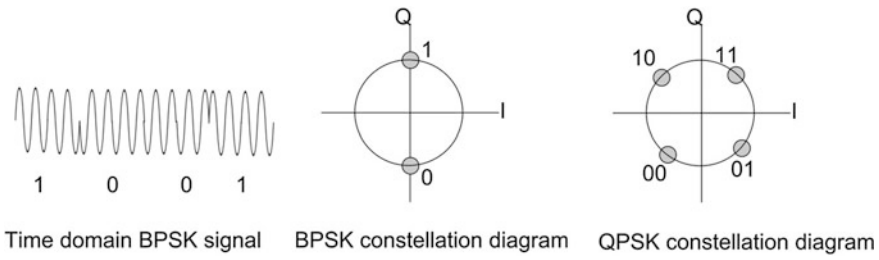


Fig. 12.4 Left shows a time-domain binary phase-shifted signal, in which the information is carried in the phase of the optical signal. Middle, the 'constellation diagram' showing the *I* and *Q* of the two symbols and their associated bits. Right, constellation diagram for four phases (quadrature phase shift keying). In quadrature phase shift keying, four discrete symbols are transmitted, each representing two bits

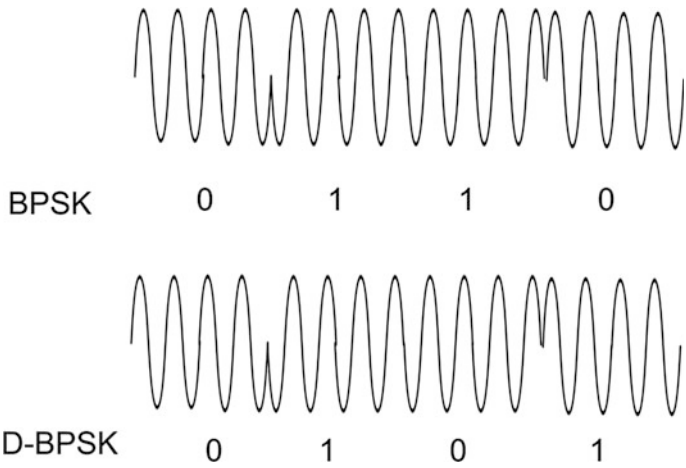


Fig. 12.5 Same time-domain bit stream, interpreted as BPSK (binary phase shift keying), where a '1' is one phase and a '0' is another phase, and differential binary phase keying (DBPSK), where a 0 is not a phase change and 1 is a phase change from the previous bit

change in phase from one bit to the next denotes a '1', while continuing the phase from one symbol to the next denotes a '0'. The idea is illustrated in Fig. 12.5, comparing phase shift keying to differential phase shift keying (Fig. 12.5).

The potential advantage is that it can be detected without the use of a local optical signal, as it can be compared to adjacent symbols to see if they are the same phase or have a phase change. The disadvantage is that receiving errors in bits propagate; an error in one bit will cause errors in subsequent bits. The mechanics of detecting optical phase changes will be discussed subsequently.

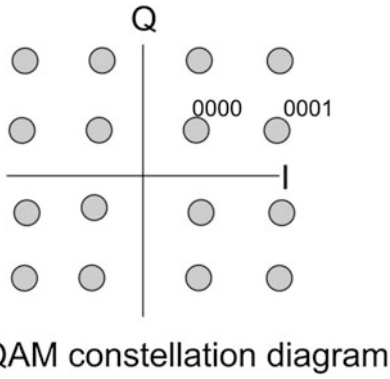


Fig. 12.6 A constellation of a 16-QAM system. In this system, both the amplitude and phase of the receiver are simultaneously modulated

12.3.3 Quadrature Amplitude Modulation

Extending the idea to both amplitude and phase modulation gives rise to the following constellation of allowed symbols. This method is called quadrature amplitude modulation (QAM); the amplitudes of the in-phase (cos) and quadrature (sin) signals are both changed, and a specific symbol is identified by a particular position in the phasor domain.

Figure 12.6 shows a number of points in the *I*–*Q* plane, each of which individually represents a set of bits. For this selection of sixteen symbols, each symbol represents four individual bits. If the receiver is good enough to resolve them under the transmission conditions, symbol bit rate can be immediately multiplied by four (bit/symbol) and achieve that multiple in data rate.

All of this depends on the electronics. To some extent, it is true that advances in laser communication over the past two decades have been driven partly by slight improvements in the lasers and enormously by extreme improvement in the electronics. Electronics which work at tens of GHz can much more easily decode, receive, and error-correct very fast signals.

Example: Sketch the time-domain signals represented by 0001 and 0000 in the diagram above, indicating the relative phase (with respect to the reference cos and amplitude with respect to each other.)

Solution: Looking at it, the reference is taken as a cosine, which for convenience we will take as amplitude 2. The 0000 point is taken as amplitude one and a phase of 45°, leading the reference. The 0001 point has an amplitude $\sqrt{1^2+2^2}$, or 2.23, and a phase of arctan 1/2, or 26°. The 0000 point has amplitude of $\sqrt{1^2+1^2}$, or 1.49, so the 0001 point has an amplitude 2.23/1.49 = 1.49.

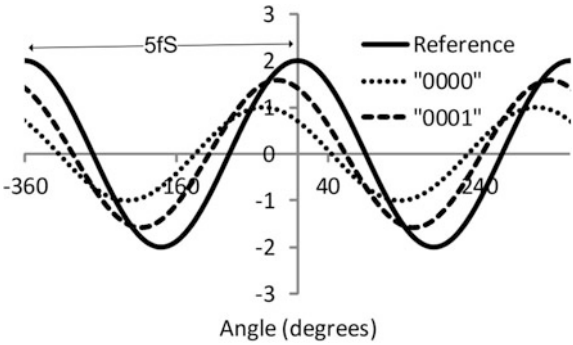


Fig. P12.7 Time-domain representation of the phasor signal

All of this information is put into Fig. P12.7. The period of the signal is indicated to reinforce that this is an optical signal.

In effect, in coherent communication the circuit is discriminating between the two lower-amplitude waves shown, at a timescale of GHz. We hope at this point the reader is comfortable relating an optical signal in the time domain to a constellation diagram with a symbol by its amplitude and phase.

This is the basic idea of coherent communication. The symbol, instead of being just an amplitude change, contains amplitude and phase. It works because the detector locks on to the particular optical frequency and excludes noise at other frequencies.

A general transmitted symbol in a coherent format can be represented as

$$L(t) = A(t)\cos(2\pi f + \theta(t)) \tag{12.5}$$

in which what is changing is both the amplitude, $A(t)$, and the phase, $\theta(t)$. These change at the symbol rate of tens of Gb/s—for example, 25 Gb/s. However, a 16-QAM system with symbols that change at 25 Gb/s is transmitting 100 Gb/s on a single laser.

With coherent systems, another multiplexing doorway becomes open to us—polarization division multiplexing.

12.3.4 Polarization Division Multiplexing

Apart from wavelength, another thing which can be easily distinguished in optical signals is polarization. Two different optical signals at orthogonal polarizations can be separately filtered at the receiving end. This additional factor of two applies to every single wavelength in the ITU Grid, and so, the bandwidth is potentially doubled.

However, standard single-mode fiber is not designed to be polarization-maintaining. While there are two orthogonal polarization modes in the fiber (one

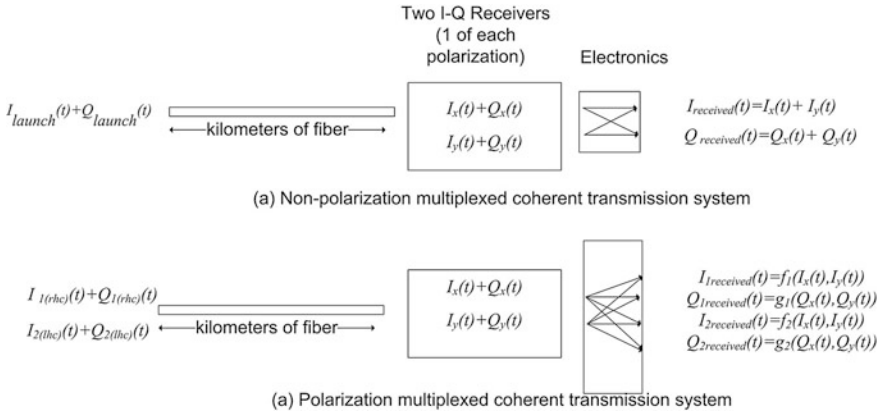


Fig. 12.8 A coherent system without polarization multiplexing and with polarization multiplexing. **a** A coherent system which does not use polarization multiplexing simply adds together the *I* and *Q* components in both polarizations to obtain the received signal; **b** a coherent system with polarization multiplexing transmits one signal on each polarization. Even though the polarizations get scrambled during transmission, the electronics can reconstruct the original transmitted bit stream based on signature bits in the bit stream or previous characterization of the link

TE-like and one TM-like), minor imperfections and stresses in the fiber as it is fabricated disturb the polarization and mix the two polarizations together. An optical signal launched with one specific polarization will become mixed with the other polarization before very long.

In a simple amplitude system, a signal launched with no specific polarization is received by a receiver that is not polarization sensitive, and the amplitude is recovered. In this situation, polarization is not relevant.

However, with polarization filter receivers and the knowledge of the entire optical field achieved in a coherent system, the signals can be reconstructed. With a coherent transmitter, it is possible to transmit polarization-multiplexed signals on standard fiber.

Figure 12.8 illustrates this idea. In a coherent system without polarization multiplexing, a signal is launched. The transmitted signal gets mixed into two polarizations, *x* and *y*. The receivers at the end recover the amplitude and phase of both polarization components and add them together.

In a coherent system with polarization multiplexing, two signals in two different polarizations are launched on a fiber. At the other end of the fiber, the received signal in one polarization contains components of both launched signals (and similarly, the received second polarization is made up partly from both of the two launched signals). The processing electronics reconstruct and separate the two orthogonally polarized bit streams based on a prior characterization of the link or signatures in the bit stream. Thus, automatically the capacity of the link is doubled.

Multiplexing signals onto a variety of wavelengths (and polarizations) lead to an astonishing potential bandwidth for optical communications. Many experiments have demonstrated transmission rates of >Tb/s over hundreds of kilometers of standard single-mode fiber using these methods.

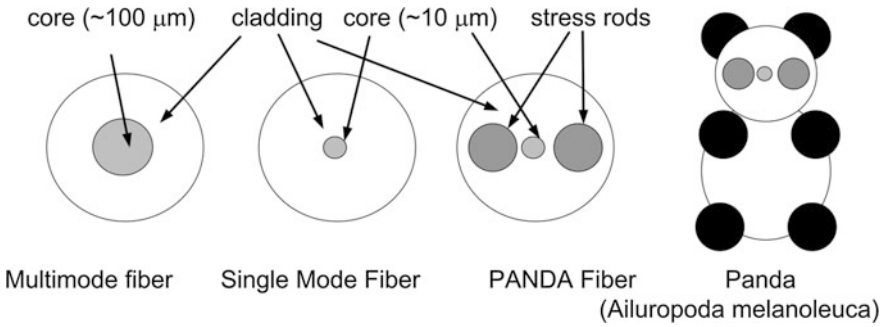


Fig. 12.9 Illustration of multimode, single-mode, and PANDA fiber

Example: Calculate the ultimate capacity of a system on the ITU 100 GHz Grid, with individual channels transmitting in a binary phase shift-keyed format at 2.5 Gbaud/s, transmitting on the 25 GHz Grid with 400 channels. This system is polarization multiplexed as well as wavelength multiplexed.

Solution: Very straightforward! 400 channels × 2.5 Gbaud/s × 2 bits/symbol (BPSK) × 2 (for polarization multiplexed), or 4 Tb/s! This is why optical communications are so valuable!

12.3.5 Polarization-Maintaining Fiber

As an aside, there are special fibers designed to maintain the polarization of launched signals. Figure 12.9 shows standard single-mode fiber, and polarization-maintaining and attenuation-reducing fiber (PANDA). The fiber is designed to maintain specific polarization coupled into it, without allowing one mode to couple into another.

These special-purpose fibers are not used for long-distance links. They are significantly more expensive than standard fiber, and their attenuation is higher. However, they do provide a direct solution to transmitting and receiving a specific polarization.

The colorful name for PANDA fiber should end all doubt as to whether optical engineers and scientists have a sense of humor.

12.4 Coherent Optical Transmitters

These coherent transmission formats have many advantages over directly modulated lasers. They can certainly be higher bandwidth generally. Because they are externally modulated, they are lower chirp and do not suffer from the inherent directly modulated laser *K*-factor limitations of resonance frequency and damping

factor. Here, we will discuss some of the basic coherent transmitters at the component and physics level.

In previous sections, we have described the transmission format in terms of the I and Q of the received optical signal. Now let us examine how we generate, and measure, I and Q.

12.4.1 Binary (or More) Phase Shift Keying Transmitter

The simplest imaginable coherent transmitter is something which alters the phase of the output based on an input voltage. Physically, it could be a lithium niobate waveguide, as discussed in Chap. 11.

The physical device is shown in Fig. 12.10. A constant wave optical signal is applied, and through the application of voltage on the high-speed electrical ports, shown, its phase is shifted.

This transmitter can realize binary (or potentially more) phase shift keying.

The voltage level sets the phase shift. With the application of V_π, the phase is shifted 180°. Commensurately lower voltage values with shift it by 90°, or 45°, or arbitrary values (however many levels the receiver is able to distinguish). Limitations to the speed are set largely by the capacitance of the electrical input, but are not related to the laser signal.

The physics of the phase shift due to the elector optic or plasma effect is described in Chap. 11.

Example: What would you see if you connected the output of the phase modulator above, to a conventional high-speed oscilloscope equipped with an optical to electrical connector, and the electrical input was driven to various phases at a symbol rate of 10 Gb/s?

Answer: After a moment's thought, you should conclude that you would see a flat line. In fact, photodetectors that do not mix signal with a reference are strictly amplitude detectors and cannot see the phase of the optical signal (which, after all, is at hundreds of THz). Detection of the phase of the optical signal and decoding the symbol require different techniques, which we will discuss in subsequent sections.

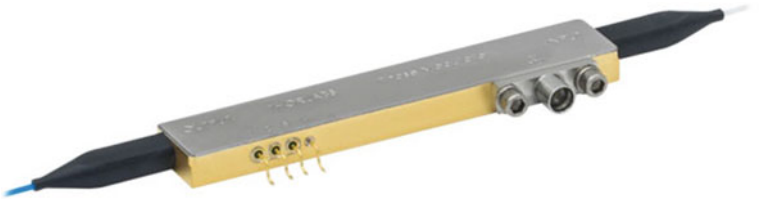


Fig. 12.10 Lithium niobate phase modulator, showing the optical input and output, the modulating electrical input and connectors for a monitoring photodiode. From Thorlabs, used by permission

12.4.2 Quadrature Amplitude Modulation

In quadrature amplitude modulation circuitry, both the in-phase and 90° out-of-phase (quadrature) parts of the signal are simultaneously amplitude modulated. A typical transmitter configuration is shown in Fig. 12.11.

This can be understood by walking through the path of the light. The laser optical carrier signal is input on the left edge and split into two, half of which will become the amplitude-modulated in-phase (*I*) signal and half of which will become the amplitude-modulated out-of-phase (*Q*) signal. The two signals then each go through a Mach–Zehnder interferometer, shifting their respective amplitudes to map to the point of the constellation shown in Fig. 12.6. One arm is shifted out of phase by 90°, typically by a 90° phase delay ($\sim$125 fs propagation delay for a 1.55 μm signal). The two signals are then recombined and transmitted.

As a phase-sensitive system, it is important that the signal remains coherent, so the source has to be of relatively narrow linewidth. In general, the laser requirements are very different than those of directly modulated lasers; apart from being low linewidth, they often have higher power requirements, as there is significant coupling loss (called, 'insertion loss') onto and out of each component.

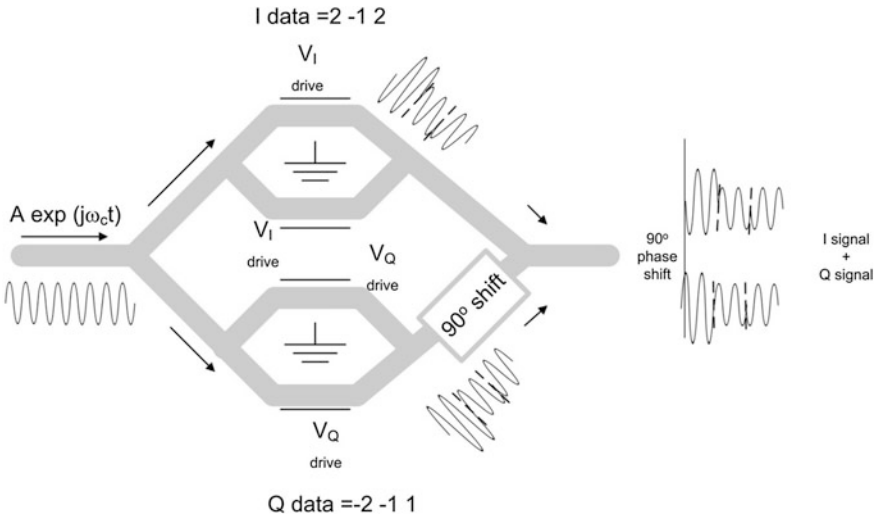


Fig. 12.11 A typical quadrature amplitude modulation transmitter. A reference signal is input into an pair of Mach–Zehnder interferometers, as shown. It is split into its *Q* half and *I* half, and both of them are modulated as required to encode the appropriate signal. The *Q* half is then shifted by 90° to put it in quadrature with the reference signal, and the two halves of the signal recombined

12.5 Receivers

The basic architecture of the transmitters is relatively straightforward. Phase shifting is done by a variety of mechanisms, and through splitting and recombining, optical signals of different amplitudes and phases can be generated. It is the electronics' job to drive all the modulators and phase shifters to generate the appropriate signals.

The receiver, however, is conceptually more complex. What is needed is an architecture, built around a combination of electronics and optics, to recover the in-phase and quadrature components of an optical system and turn them back into conventional 1's and 0's. The complete phase and amplitude can be recovered through interference with a reference signal both in phase and in quadrature.

In the section below, we outline the building blocks of these coherent receivers.

12.5.1 Reference Signal

The key to decoding the phase of an optical signal is interference with a reference signal, at the same optical frequency as the transmitted signal. That is not a trivial task, but it is made easier these days by the use of high-frequency electronics. Generally, the system starts with a local laser of the same channel as the incoming light. Table 11.1 shows that the signals are spaced by 100 (or 25) GHz. Different lasers nominally on the same channel could potentially be off by something around that. Interference with a reference laser at the frequency would produce, instead of DC signals, signals which vary at that difference frequency.

It is possible to incorporate a circuit that will phase lock the reference laser to an incoming signal via a phase lock loop. By feedback into the laser, the lasing frequency can be slightly changed until it matches exactly with the input signal as shown in Fig. 12.12.

In many modern systems, however, the electronics can actually live with GHz signal difference between the reference oscillator and the optical signal. Such systems are called intradyne systems, where the frequency difference between the reference and the signal is large by electronic standards (GHz) but not by optical standards.

In Fig. 12.12, both a more complex phase-locked reference (which is generally, kept almost exactly at the wavelength of the incoming signal) and a simpler intradyne system are illustrated. The system on the left produces a laser that is locked to the input and can be used as a reference. The system on the right simply uses a local reference laser that is close to the input signal and produces an output that is not DC, but rather oscillates at the difference frequency of the order of GHz.

12.5.2 Balanced Photodiode

A key component to receiving the signals is the mixer, pictured in Fig. 12.13: a way to combine the optical signal with the reference and obtain the difference frequency.

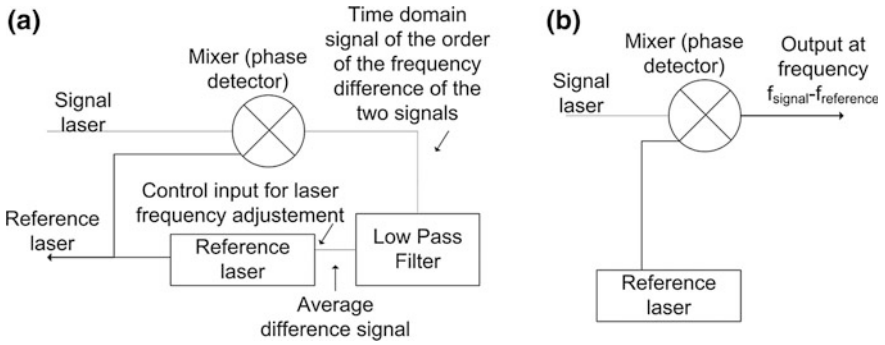


Fig. 12.12 **a** Locking to an input signal laser to create a reference signal exactly in phase with the input, **b** operating in an open-loop, intradyne mode, in which a local laser that is 'close enough' to the input signal is used, and the output oscillates at the difference frequency

One common way is a balanced photodiode pair. A typical architecture is shown below. In this discussion, the local reference wavelength is assumed to be exactly matched to the signal wavelength.

The incoming signal is combined with the reference laser and used to illuminate a photodiode. For example, let us look at the top photodiode. The photodiode is fast enough to response to $I(t)$ (the symbol rate—about 25 Gb/s) but not fast enough to respond to an optical signal, ω_c. The incoming optical signal is

$$\text{Input}(t) = I(t)\cos\omega t + Q(t)\sin\omega t + A\cos\omega t \tag{12.6}$$

where ω is the optical frequency, $I(t)$ and $Q(t)$ are the in-phase and quadrature envelopes of the signal, and A is the amplitude of the reference lasers. The output seen by the photodiode will be the time average of the signal over the timescale that it can respond to, or

$$\begin{aligned}\text{Photodiode signal}(t) &= \langle I(t)\cos\omega t + Q(t)\sin\omega t + A\cos\omega t\rangle \\ &= \frac{1}{2}\left(I^2(t) + Q^2(t) + AI(t) + A^2(t)\right)\end{aligned} \tag{12.7}$$

consisting of two DC offsets and a mixed signal, $AI(t)$, which directly contains the transmitted information at the symbol rate. The I and Q vary as the symbol rate and can be detected by the photodetector.

This arrangement has two specific problems. First, it can be difficult to pick out the signal from among the other terms. Second, the signal $I(t)$ is often quite weak, and it is a question of signal-to-noise ratio in discerning it out from the incoming information. A clever arrangement called a balanced photodiode pair can address both of these problems.

In a balanced photodiode, there is a second photodiode (shown at the bottom of Fig. 12.13). The arrangement of the second set of photodiodes is the same as at the top, except that the reference laser has its phase shifted by 180°, inverting it. This makes the input

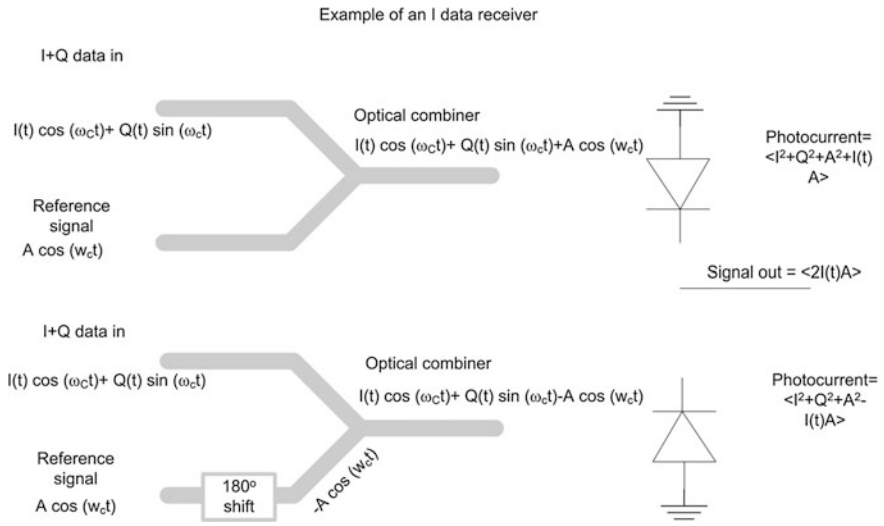


Fig. 12.13 A balanced photodiode architecture, in this case for detecting the '*I*' component. The signal (containing both *Q* and *I*, and the reference local laser) are combined and illuminated on a photodiode. The photodiode gives the response shown, proportional to the amplitudes of the individual components and the mixed signal. By inverting the reference, the difference signal between the two photodiodes can be made directly proportional to *I*. To obtain the *Q*, the second pair of photodiodes is used, with the reference signal shifted by 90° before mixing

$$\text{Input}(t) = I(t)\cos\omega t + Q(t)\sin\omega t - A\cos\omega t \tag{12.8}$$

and the output

$$\begin{aligned}\text{Photodiode signal}(t) &= \langle I(t)\cos\omega t + Q(t)\sin\omega t - A\cos\omega t\rangle \\ &= \frac{1}{2}(I^2(t) + Q^2(t) - AI(t) + A^2)\end{aligned} \tag{12.9}$$

Now, notice what happens if the difference of the two photocurrent signals is taken, as if they were across the same resistor in different directions. The difference is

$$\begin{aligned}\text{Photodiode difference signal}(t) &= \frac{1}{2}((I^2(t) + Q^2(t) + AI(t) + A^2) \\ &\quad - (I^2(t) + Q^2(t) - AI(t) + A^2)) \\ &= AI(t)\end{aligned} \tag{12.10}$$

The common terms have been eliminated, and the term of significance $AI(t)$ has been doubled.

In addition, this particular arrangement of photodiodes has the intriguing feature that the signal to noise of the actual signal $AI(t)$ can be increased simply by increasing the reference laser power A.

The figure shown will obtain the value $AI(t)$; a slightly modified circuit will obtain the $AQ(t)$ component. Having decoded the $I(t)$ and $Q(t)$, the electronics determine the position in the constellation for quadrature encoded systems.

12.5.3 A Full Coherent System

To tie all of these topics together, Figs. 12.14 and 12.15 show block diagrams of a full coherent transmitter and receiver. This block diagram is meant to give a picture

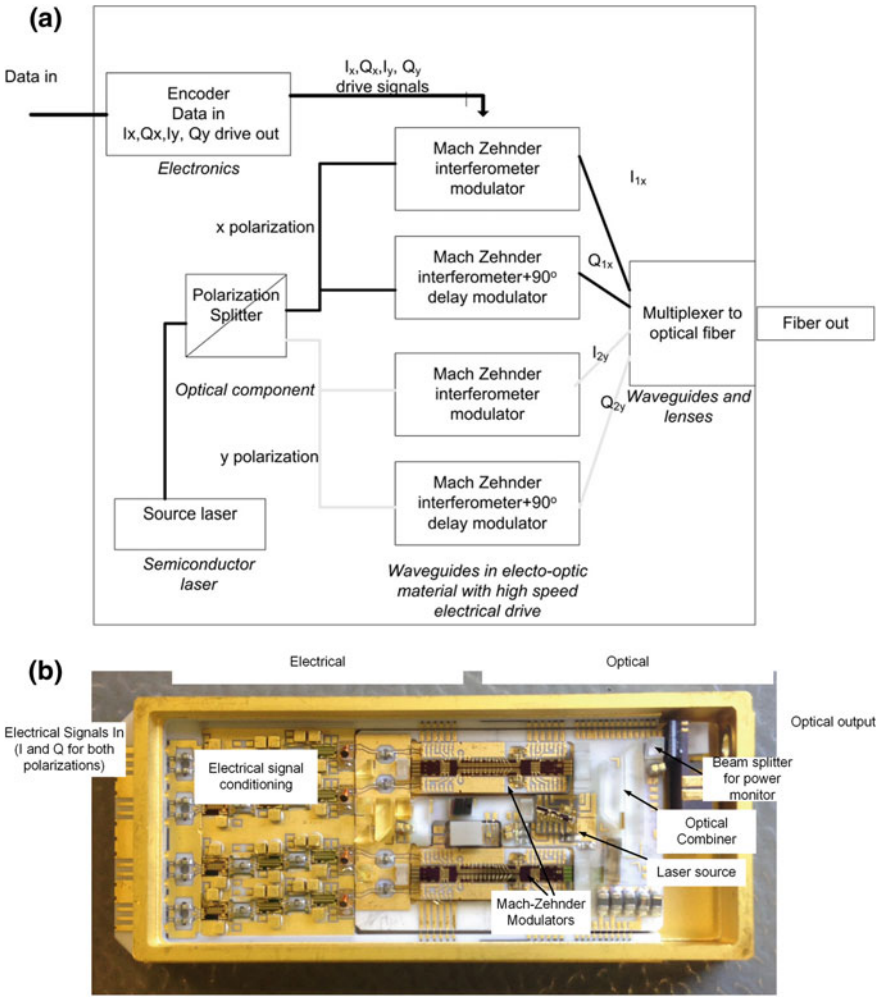


Fig. 12.14 **a** Diagram of an entire polarization multiplexed quadrature amplitude-modulated communication system. The reference laser is split into two orthogonal polarizations, x and y, and each one is separately modulated with its I and Q components. The full signal, with its I and Q components of both polarizations, is multiplexed onto a fiber and transmitted; **b** Picture of a polarization multiplexed transmitter, showing how the components are integrated and packaged

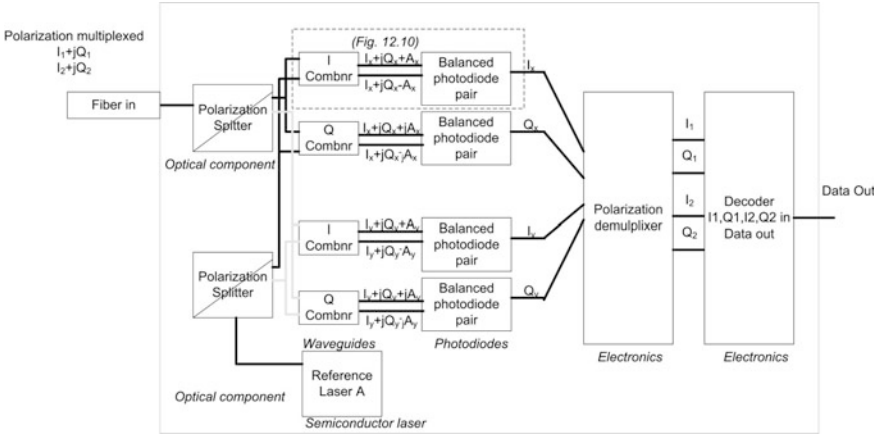


Fig. 12.15 Block diagram of a polarization division-multiplexed quadrature amplitude modulation receiver. The signal is received and split into two orthogonal polarizations which are mixed with two orthogonal polarizations of the reference laser *A*. The *Q* mixers shift the reference signal by 90°

of the information flow and identify at which points the signal is optical, where it is electrical, and again point out the role of the electronics in transmitting and receiving coherent signals.

The details of most of the boxes have been previously discussed. The laser is a semiconductor lasers, designed for low linewidth. The Mach–Zehnder interferometers are shown as block in Fig. 12.14, with two sets, on for each polarization. The electronics play a major role: The data to be transmitted is input and encoded using the chosen format (quadrature phase shift keyed, quadrature amplitude modulated, or whatnot) into the drive signals on the modulators. Two modulators are used for each polarization division-multiplexed coherent transmitter.

A similar full diagram of the receiver is shown in Fig. 12.15.

12.6 Coherent Transmission in Context

12.6.1 Comparison of Coherent and Incoherent (Amplitude Shift Keying) Systems

Let us spend a few paragraphs discussing coherent and incoherent amplitude-shift-keyed transmission in context. The incoherent system consists of a directly modulated laser diode or other amplitude-modulated light sources on the transmission end, and a receiver with a bandwidth of roughly the transmission frequency on the other end. This system is sensitive only to the optical amplitude and is straightforward and simple.

It suffers from a few disadvantages, particularly for devices based on directly modulated lasers. The laser modulation speed has a limit based on capacitance, K-factor, and other laser intrinsics. The second issue is dispersion. As the laser is modulated, the 1s and 0s travel at different speeds, leading to dispersion over a length of fiber. Externally modulated amplitude shift devices, such as Mach–Zehnder interferometers or electroabsorption modulators, have much better external speed limitations and reduced dispersion effects.

In general, amplitude shift keying [both on–off modulation and multilevel modulated (pulse amplitude modulation)] tends to be used in data centers, where cost matters. Amplitude-modulated receivers and transmitters are much less expensive than coherent systems. These amplitude-modulated devices are often used with 1310 nm devices in which dispersion is important, and loss is less important. They are also widely used in short-haul and metro links at 1550 nm wavelengths.

Coherent transmitters, on the other hand, have almost completely taken over long-haul (100 km+) transmission systems. As it happens, by completely recovering the amplitude and phase of the received signal, it is possible for the electronics to recover from nearly any amount of dispersion. Hence, dispersion compensation is not a factor for coherent transmission.

Coherent transmission also offers bit rates much higher than symbol rates and improves the bit error rate for a similar signal-to-noise ratio. In a way similar to frequency modulated radio transmission, at similar power levels, transmission is much clearer for coherent systems.

12.6.2 Communication Formats

'The nice thing about standards is that you have so many to choose from.'[1] Nowhere is that more correct than in the variety of formats usable for coherent communications. In Table 12.1 we list some (not all) acronyms for common communication formats and their decriptions.

The primary advantages to coherent communication are much improved fidelity at the same signal-to-noise ratios and transmission unlimited by dispersion.

12.7 Limits to Transmission Distance in Optical Systems

In any transmission system, optical or otherwise, the limit to getting clear transmission with no errors is determined by the signal-to-noise ratio. The signal (a '1' or a '0') competes with a background of random noise from many sources and has to stand out from the noise by a specific amount in order to be understood. The

[1]Andrew S. Tanenbaum, Computer Scientist. Another area with lots of standards!

Table 12.1 A list of optical communication formats

Format	Acronym	Description
Amplitude shift keying, non-return to zero	ASK-NRZ	Shifts in optical amplitude encode the symbols; each symbol period is one bit period
Amplitude shift keying, return to zero	ASK-RZ	Shifts in optical amplitude encode the symbols; after each bit, the symbol returns to zero
Pulse amplitude modulation	PAM	Optical amplitude is modulated over several levels; each level encodes more than one bit
Parallel series modulation	PSM	A form of wavelength division modulation in which each of typically four channels around 1310 nm is separately modulated
Binary phase shift keying	BPSK	The amplitude of the laser is kept constant, but the optical signal is shifted between one of two phases as the symbol change
Differential binary phase shift keying	DBPSK	The phase of the signal is the symbol change; however, a change of phase represents a one value and an unchanged phase the other bit value. This may obviate the need for a reference signal
Quadrature (or more) phase shift keying	Q (or more) PSK	The phase of the laser signal is shifted between several different values, each one encoding different bits of information
Quadrature amplitude modulation	QAM	Both the phase and amplitude of the signal laser are modulated over many points in optical phase plane. This is commonly done in 16, or 64, or even more points, such that each symbol represents 4, 6 or more bits respectively

amplitude of the signal itself is not that significant; you can generally amplify the optical, or electrical, signal freely. But it does little good if it sits on a bed of static.

A good analogy is how loud someone needs to shout to talk to you across a room. If the room is empty, and the noise is low, the person can speak softly and be understood. If one is at a party, the signal (voice) level needs to be raised for intelligible conversation. Similarly, signal-to-noise ratio is the key component for error-free transmission, not the amplitude.

12.7.1 Optical Signal-to-Noise Ratio

A picture of the optical signal-to-noise ratio (OSNR) is shown in Fig 12.16. In optical signals, typically the carrier frequency (of the light) is enormously larger than that of the signal. The optical signal-to-noise ratio is defined is the power in the optical signal divided by the noise in a particular bandwidth.

Mathematically, the OSNR is defined as

$$\text{OSNR} = 10\log_{10}(S/N) \tag{12.11}$$

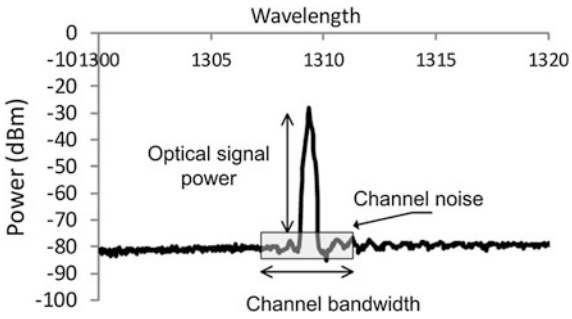


Fig. 12.16 Optical signal-to-noise ratio is defined as the optical power of the signal, divided by the noise in a filtered bandwidth. Because the carrier bandwidth is so high (THz), it is more relevant to look at the noise in an optical channel of narrow defined bandwidth

where S is the signal power, and N is the total noise power in a single channel. The bit error rate associated with the OSNR, measured in this way, is (empirically)

$$\log_{10}(\text{BER}) = 10.7 - 1.45\text{OSRN} \tag{12.12}$$

The noise power N, in particular, depends on the width of the region defined as the channel bandwidth. Typically, a number like 0.1 nm is used.

Defined in this way, the optical signal-to-noise ratio can be measured on an optical spectrum analyzer, quantifying the power in the signal and the variation in the noise around it.

12.7.2 Eye Diagram-Based Signal-to-Noise Ratio

In addition to the optical signal-to-noise ratio defined above, there is another definition of signal-to-noise ratio, SNR, appropriate to an amplitude-modulated system defined as

$$\text{SNR} = \frac{P_{\text{signal}}}{P_{\text{noise}}} = \frac{L_1 + L_0}{\sigma_1 + \sigma_2} \tag{12.13}$$

An illustration of the various quantities given is shown in Fig. 12.17.

The first definition (of Eq. 12.11) says that the signal-to-noise ratio is the signal power divided by the noise power in a channel bandwidth. Essentially, it is a frequency-domain definition, with the frequency mapped out by the optical wavelength.

Here, signal-to-noise ratio is defined as the average power level of the '1' and '0' states, $(L_1 + L_2)/2$, divided by the variation in average power of the '1' and the '0' states, $(\sigma_1 + \sigma_2)/2$. This puts it into the domain of something that can be easily seen in a laboratory in a time-domain eye pattern. When observing a pattern of optical

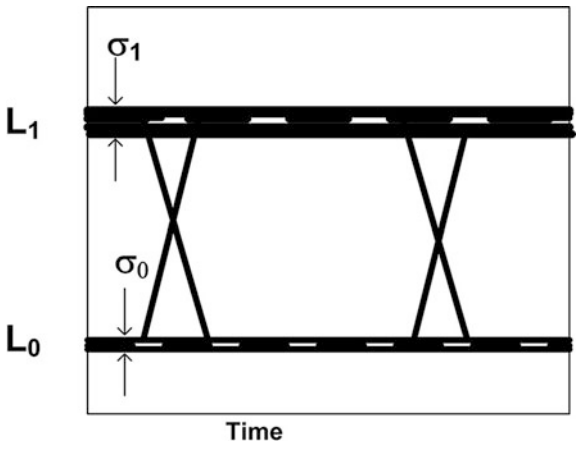


Fig. 12.17 An illustration of a noisy eye pattern, in which variations in the high and low levels are characterized by standard deviations σ_1 and σ_2, respectively. The dotted line shows the nominal L_1 and L_0 levels

pulses by shining the light on a photoreceiver, with the receiver connected to an oscilloscope, the levels are not completely flat (for a real example, see Fig. 8.1). The variation in the '1' or the '0' levels comes from noise in the light source, or the photodiode, and fundamentally limit how much information can be transmitted.

Often in amplitude-modulated formats, the high signal level is much higher than the low signal level, and so, the approximation below can be made.

$$\text{SNR} \approx \frac{L_1}{\sigma_1} \tag{12.14}$$

12.7.3 Bit Error Rate Versus Transmission Format and Signal-to-Noise Ratio

However, neither of those definitions of optical signal-to-noise ratio (Eqs. 12.13 and 12.11) take into account differences in the efficiency between various communication systems, particularly for ratios of bits to symbols. A quadrature amplitude system can transmit two, four, or six bits for a given power level while a single on–off keying system will only transmit one bit. However, the raw signal power level and the noise level can be the same.

To better compare different systems, let us introduce a slightly different noise figure. The symbol R is the ratio of the bit energy E_b to noise spectral density, N_o in W/Hz, as

$$R = \frac{E_\text{b}}{N_\text{o}}, \tag{12.15}$$

Noise spectral density is total noise power, in W, divided by channel bandwidth in Hz; hence, the units of it are joules as well.

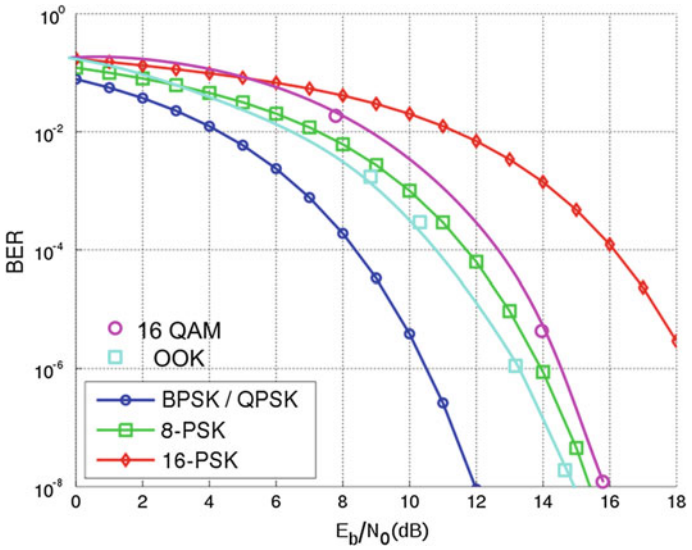


Fig. 12.18 Bit error rate versus bit energy-to-noise ratio for various detection systems. After Wikipedia, https://en.wikipedia.org/wiki/Eb%2FN0

Because R is energy/bit, not energy/symbol, it is a better metric for comparing transmission fidelity at different signal-to-noise ratios across coherent and incoherent formats.

Figure 12.18 shows bit error rate as a function of R for different formats.

To illustrate how this is used, on the given chart, the y-axis shows bit error rate, and the x-axis shows the ratio R. For a given ratio R (e.g., 10 dB), a BPSK transmission format would give a bit error rate of 10^{-5}–10^{-6}, while amplitude shift keying (on–off keying) would give a bit error rate of 10^{-3}–10^{-4}. In general techniques which have lower bits per symbol have better bit error rates for a given R. The major tradeoff is in the extra complexity of the coherent transmitter and receiver.

The quantity R is not fundamentally different than OSNR or SNR, as shown in the example below.

Example: A QAM-16 system has an OSNR of 48. What is R, the ratio of bit energy/noise spectral density? (assume the channel bandwidth is $1/\tau$, where τ is the symbol period).

Solution: This question asks to convert signal power and noise power, into energy per bit. Let us assume a symbol period τ and an average power level S. The total energy per symbol is τS, and hence, the total energy per bit is $\tau S/4$ (since

QAM-16 has four bits/symbol). The total noise power N converts to noise spectral density by dividing it by $1/\tau$. Hence

$$\frac{E_b}{N_o} = \frac{S\tau/4}{N/1/\tau} = \frac{S}{N4}$$

or 12.

Equations 12.13 and 12.15 generally differ only by the factor of bits/symbol.

12.8 Noise Sources

Noise is generated throughout an optical transmission system, by the laser, at the receiver, and even in the fiber and amplifiers. All of this noise figures into the noise component seen by the receiver.

Figure 12.19 lists some of the noise sources which appear in optoelectronic transmission links.

The diagram lists the sources of noise, and where they occur. The ones listed in black type will be discussed in some detail with the ones listed in gray will be roughly defined.

One note about expressions that appear for noise in many of these terms. In optical systems in particular, the carrier frequency of hundreds of THz is much, much larger than the signal bandwidth, of perhaps GHz. It makes no sense to look at total noise in the carrier signal when the receivers are filtered to look at a particular signal band. Hence, the expressions for noise all carry within them a bandwidth B, which is equivalent to the bandwidth of the channel being received. For example, in a WDM system spaced by 25 GHz, the relevant noise would only be within that 25 GHz bandwidth.

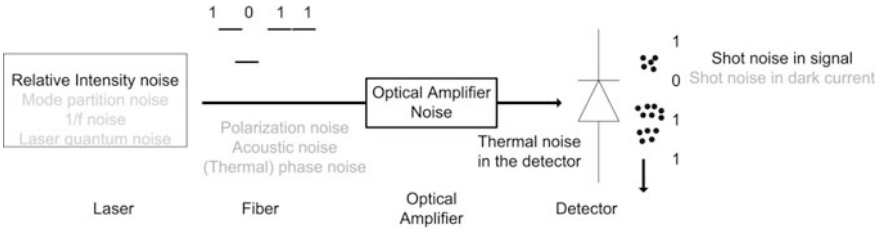


Fig. 12.19 Noise sources in an optoelectronic communication system. The ones in dark type will be discussed in detail below, and the ones in gray only mentioned and defined

12.8.1 Relative Intensity Noise

In Chap. 8, we introduced relative intensity noise. This is noise inherently associated with the laser, due to amplified spontaneous emission. Generally, commercial lasers have a specification for relative intensity noise of average levels ~ −145 dB/Hz, and here, we would like to specify the impact it has on transmission.

Relative intensity noise is essentially intensity fluctuations in the laser output, and it increases the variation associated with the laser signal above. The relevant formula is

$$\sigma_{\text{RIN}} = L\sqrt{\text{RIN} * B} \tag{12.16}$$

in which relative intensity noise spectral density (usually in dbW/Hz) should be written as a ratio, B is the channel bandwidth, and L is the optical power.

Example: What is the noise and signal-to-noise ratio associated with a 1 μW signal in a laser with a typical RIN of −140 dB/Hz and a 25 GB/s bit rate?

Solution: The RIN, written as a ratio, is $-140 = 10\log_{10}(\text{RIN}_{\text{W/Hz}})$, or $\text{RIN}_{\text{W/Hz}} = 10^{-140/10} = 10^{-14}$ Then $\sigma = 1\,\mu\text{W}\sqrt{25 \times 10^9 * 10^{-14}} = 15\,\text{nW}$ This signal-to-noise ratio of roughly 60 dB is not a very big limitation.

12.8.2 Shot Noise

A second important source of noise in optical communication is what is called shot noise. Shot noise comes about because, as shall be shown, optical signals transmitted for a long time can have a very low power level (e.g., μW), and each bit can occupy a remarkably short time. A bit in a 25 Gb/s signal lasts just 40 ps. A quick calculation shows this signal has an astonishing 25 photons per bit (or, looked at from the receiver point of view, involves detecting 25 photons).

A typical model for a distribution of 25 generated photons, each of which independently faces the same probability of generation and detection, is a Poisson distribution. In such a distribution, the standard deviation σ_{sd} of a population of N independent events is equal to

$$\sigma_{\text{sd}} = \sqrt{N} \tag{12.17}$$

(with N here the number of photons). This variation from the nominal average leads to noise in the signal. This signal is seen by the photodetector in Amps, not Watts. The expression below for photoreceiver shot noise in A is

$$\sigma_{\text{shotnoise-Amps}} = \sqrt{2B\eta Lq} \tag{12.18}$$

where B again is channel bandwidth, L is the signal level (in W), and q is the electronic charge. The symbol η is the responsivity of the photoreceiver, in A/W. To get the noise in W references back to the optical signal, the photodiode noise is divided by the responsivity, as

$$\sigma_{\text{shotnoise-W}} = \frac{1}{\eta}\sqrt{2B\eta Lq} \tag{12.19}$$

Example: What is the shot noise (in W, and A), of a 1 μW 10 Gb/s signal impinging on a photoreceiver with responsivity of 0.8 A/W?

Solution: The shot noise in A is $\sqrt{2 * 10^{10} * 0.8 * 10^{-6} * 1.6 * 10^{-19}} = 5.5 * 10^{-8}$ A and in W, $5.5 * 10^{-8}/0.8 = 69$ nW.

12.8.3 Erbium-Doped Fiber Amplifier Noise

In any long-distance system, there are erbium-doped fiber amplifier (EDFA) links. As we talked about back in Chap. 1, the EDFA takes 1 μm light and pumps a population of erbium atoms into inversion, such that they create an optical gain medium for 1.55 μm light. Such amplifiers typically amplify light by 30 dB or more and are one of the key enabling blocks for optical transmission.

Ideally, the amplifier would amplify both the incoming signal and noise, and the signal-to-noise ratio would remain unchanged. But in fact, through a process called quantum noise, the EDFAs themselves actually add noise to a signal. Though a detailed description is beyond the scope of the book, we will give a qualitative overview.

The noise in this process comes from a few factors. The erbium gain rides on a background of spontaneous emission in the amplifier from the erbium level. This background is also amplified and added to the output signal. This is thought to be the dominant noise source in EDFAs.

In addition, the 1 μm pump laser has noise of its own, which translates into 'gain noise' in the amplifier. Also, as the laser signal grows stronger as it is propagated though the amplifying fiber, some of the signal photons can be reabsorbed. The absorption process is statistical in nature and carries with it a shot noise term. If a nominal number P of photons are absorbed or emitted, these photons follow a Poisson distribution with a variance of P and standard deviation of $\sqrt{P}$. This statistical variation translates into additional noise on the output.

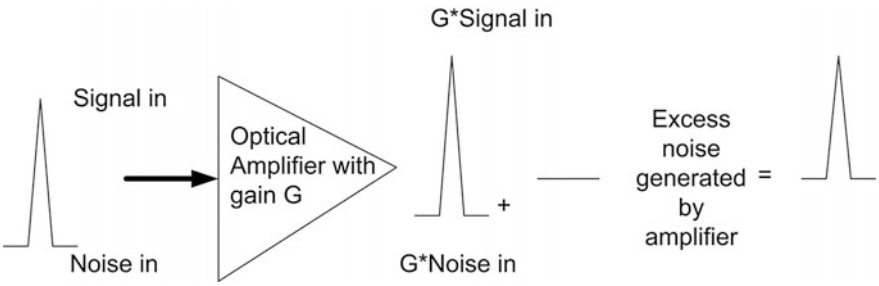


Fig. 12.20 An illustration of the noise added by an erbium-doped fiber amplifier

This increase in noise is characterized by a noise factor F or a noise figure NF, which is defined as

$$F = \frac{\text{SNR}_{\text{in}}}{\text{SNR}_{\text{out}}} \tag{12.20}$$

$$\text{NF} = 10\log(F) = 10\log(\text{SNR}_{\text{in}}) - 10\log(\text{SNR}_{\text{out}}) \tag{12.21}$$

Even though both the noise and the signal are multiplied by the amplifier gain G, the amplifier itself adds noise. The signal-to-noise ratio after amplification is lower than the signal-to-noise ratio before amplification. This is illustrated in Fig. 12.20.

Noise figures for ideal erbium-doped fiber amplifiers are always greater than 3 dB and are typically 5–6 dB.

Example: An optical signal of 10 μW and a signal-to-noise ratio of 20 enters an EFDA with a gain of 30 dB and a noise figure of 6 dB. Calculate the exit signal power, exit noise power, and exiting signal-to-noise ratio.

Solution: A gain of 30 dB means a gain of a factor of a 1000, as 30 = 10 log G, or G = 1000. If the input power is 10 μW, the output power is then 10 mW. The initial signal-to-noise ratio is 20 (or, 10 log (20) = 13 dB) with an initial noise power of 500 nW within the determined bandwidth. To calculate the output noise power, we use the noise factor. If the noise factor is 6 dB with an initial noise factor of 13, by Eq. 12.21, then the signal-to-noise ratio is 7 dB after the amplifier. Hence, 7 = 10 log (10 mW/N), or N = 2 mW. The noise increased by 36 dB while the signal power increase by 30 dB, for a consequent degradation of 6 dB in the SNR.

This example illustrates that amplifiers alone are not the solution to transmission for infinite distances. Ultimately, it the signal-to-noise ratio which fundamentally limits transmission distances.

12.8.4 Thermal Johnson Noise

Apart from the variance in the incoming signal, the receiver photodiode has noise associated with it even without an impinging signal. Such noise, called Johnson noise, comes from the thermal random motion of the carriers, as shown in Fig. 12.21.

These randomly moving carriers cross an equivalent resistance, R, and give rise to a Johnson noise (in V) of

$$\sigma_{\text{Johnson}} = \sqrt{4BkTR} \tag{12.22}$$

In which B is channel bandwidth, k is Boltzmann's constant and T is the absolute temperature. The symbol R is the equivalent circuit resistance seen by the photodiode and depends on the circuit in which it is placed.

At the end, we want to be able to add the noise due to the various causes, so we need to reference this voltage noise back to current noise and then to optical signal noise.

$$\sigma_{\text{Johnson}-\text{Amps}} = \frac{1}{\text{Gain}}\sqrt{4BkTR} \tag{12.23}$$

in which the gain is the circuit gain, in V/A (often the feedback resistance of the transimpedance amplifier). Referenced back to optical signal, the equivalent Johnson noise of the optical signal is

$$\sigma_{\text{Johnson}-\text{optical}} = \frac{1}{\eta}\frac{1}{\text{Gain}}\sqrt{4BkTR} \tag{12.24}$$

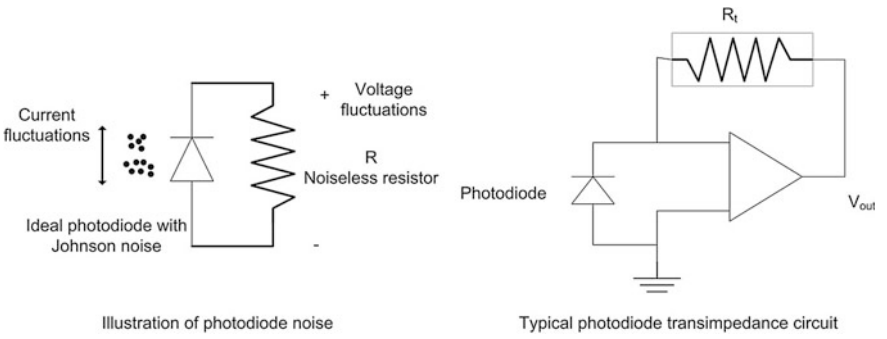


Fig. 12.21 Left: illustration of thermal Johnson noise on a diode detector. The temperature gives rise to random carrier motion in the diode, which appear as a voltage across the equivalent circuit resistance. Right, illustration of how the current motion leads to voltage noise in a typical circuit

This noise is present whether or not there is a signal and is independent of the signal magnitude.

Example: At room temperature, a photodetector is used to detect a 25 Gb/s signal. The photodetector has a 10 K resistor in the circuit to convert the photocurrent into a voltage and a responsivity of 0.8 A/W. Find the equivalent Johnson optical noise

Solution:

$$\sigma_{\text{Johnson-volts}} = \sqrt{4 * 25 * 10^9 * 300 * 1.38 * 10^{-23} * 10{,}000} = 6.5 * 10^{-4}\,\text{V}$$

To convert to Amps, the gain is needed; as the photocurrent goes across the resistor of 10,000 Ω, the gain is 10,000 V/A.

$$\sigma_{\text{Johnson-Amps}} = \frac{1}{10{,}000} 6.5 * 10^{-4}\,\text{V} = 6.5 * 10^{-7}\,\text{A}$$

And then to optical Watts, is

$$\sigma_{\text{Johnson-Watts}} = \frac{1}{0.8} 6.5 * 10^{-8} = 81\,\text{nW}.$$

12.8.5 Combination of Noise Sources

The noise sources, each essentially random and uncorrelated, are combined in quadrature to get the total noise.

$$\sigma^2_{\text{noise}} = \sigma^2_{\text{shotnoise}} + \sigma^2_{\text{Johnonnoise}} + \sigma^2_{\text{RIN}} + \cdots \tag{12.25}$$

From the total noise, the ratio of signal-to-noise (Q) can be computed, and the bit error rate can be estimated.

Example: For the three previous examples, we calculated values of shot noise of 69 nW, Johnson noise of 81 nW, and laser relative intensity noise of 15 nW for an input signal with an input power of 1 μW. From these three, calculate a signal-to-noise ratio (SNR), and using

Table 12.2 List of other sources of optical system transmission noise

Laser Noise Sources	Explanation
1/*f* noise (flicker noise)	Low-frequency noise source of indeterminate cause, but generally not significant at high (>Gbaud/s) symbol rates
Mode partition noise	Noise related to having more than one optical mode in a device. Typically associated with multimode Fabry-Perot lasers
Laser quantum noise	Noise associated with the Poisson distribution of the number of photons emitted
Fiber noise sources	
Polarization noise	Related to coupling between modes of different polarization in the fiber
Thermal noise	Noise related to temperature changes on the fiber
Acoustic noise	Noise related to fiber vibration and associated stress/index changes

Fig. 12.18, estimate the bit error rate for an amplitude shift keyed (on–off keyed) transmission format.

Solution: The three noises add in quadrature as $\sigma^2_{\text{noise}} = (81\,\text{nW})^2 + (15\,\text{nW})^2 + (69\,\text{nW})^2 = (107\,\text{nW})^2$, so the total noise power is 107 nW.
The signal-to-noise power ratio, Q, is 1000/107 = 10. This gives roughly a bit error rate of 10^{-3}. Notice, of course, that binary phase shift keying would lead to a much better bit error rate of 10^{-6}.

12.8.6 Other Noise Sources

Listed in Fig. 12.19, above, are other noise sources (and there are of course others that can be imagined). Table 12.2 briefly defines those noise terms.

All of these noise sources cooperate to hide the signal, and it is the ratio of the signal strength to the total noise which governs the detectability and speed of the system.

12.9 Final Words

Here, we come to the end of the book, but not, fortunately, to the end of the subject. There are many fascinating topics in the broad area of semiconductor lasers that we have not even touched upon.

We have focused in this book on topics that concern communications lasers at the conventional 1.3 or 1.55 μm wavelength. The factors which control directly modulated speed have been covered and the reader should come away with a good understand of the limitations and capabilities of directly modulated, distributed feedback lasers.

We have also talked about external modulation techniques, including integrated laser modulators and Mach-Zehnder modulators. The highest performance optical transmission systems do not use direct modulation; they use external modulation, which is typically combined with techniques for coherent transmission and forward error correction. Performance from these systems is quite incredible. In October 2018, Fujitsu demonstrated 600 Gb/s transmission on a single wavelength using 64-QAM. Commercial switch systems up to 500 Gb/s are available. There is hardly any limit to the bandwidth available over a fiber.

All of this capability is built upon semiconductor lasers and imposes stringent requirements upon the lasers. We hope that with the aid of this book, these laser requirements can now be appreciated and (if this is your job function) satisfied.

There are also some fascinating new areas in laser materials, all invented since the beginning of the 1990s. The development of high-efficiency blue LEDS and *blue lasers* based on GaN on sapphire was a phenomenal breakthrough, enabling new applications for displays and for solid-state lighting using shorter wavelength lasers. On the very long-wavelength side, a team at Bell Laboratories developed a method to use conventional semiconductors, with bandgaps around 1 eV or higher, to emit very low energy and very long-wavelength photons. The *quantum cascade laser* is now widely used in spectroscopy and is the most convenient method for the generation of long-wavelength sources.

The first semiconductor laser was demonstrated using bulk semiconductors at low temperature, but quantum wells have been the standard material for semiconductor lasers for many years. The extra confinement they provide compared to bulk material allows for good performance and room temperature or higher operation. However, recently, practical *quantum dot materials* have emerged. These materials have demonstrated lower threshold current density and higher temperature independence than any quantum well device. Quantum dot active regions are currently being developed as a potential alternative to quantum well active regions for applications in optical communication and other areas.

12.10 Summary and Learning Points

In this chapter, we have looked at coherent communication systems which make up the bulk of long distance, high speed links. The formats used for encoding bits with phase and amplitude were described, and the methods for generating and detecting coherent signals were outlined. We also discussed sources of noise in laser communication systems and how they limit transmission.

A. Coherent light can be represented as a phasor, with the amplitude and phase with respect to a reference shown as a vector on a two-dimensional plane.
B. Coherent transmission methods modulate and detect both the amplitude and phase of the light and encode and decode that light into data.
C. Coherent transmission has many different formats, with various degrees of complexity and bits/symbol associated with them.

D. Binary phase shift keying (BPSK) shifts the phase of the optical wave by 180° to signify one symbol or another, while keeping the amplitude of the wave the same.
E. Quaternary phase shift keying (QPSK) shifts the phase by 90°, while keeping the amplitude of the wave the same. It can be extended to eight, sixteen, or arbitrary phase levels, depending on the sensitivity of the receiver.
F. In quadrature amplitude modulation, both the phase and amplitude of the wave are modulated. This is usually named by the number of distinct symbols that are created: 16 QAM, 64 QAM, and so forth.
G. These systems have the advantages of lower symbols rates for a given bit rate compared to directly modulated systems. They can also operate at lower signal-to-noise ratios than amplitude-only systems.
H. These transmission formats need high-speed electronics to receive and recover the signal.
I. Coherent systems can electronically compensate for polarization mixing (leading to polarization division multiplexing on standard fiber) and dispersion.
J. The mechanics of the phase of the optoelectronics transmitters involve phase shifts with optoelectronic materials (typically lithium niobate, via the electrooptic effect, or silicon waveguides, via the plasma effect).
K. Phase shifters by themselves do the phase shifting required for phase shift keying and create the in-phase and quadrature components for coherent transmitters.
L. A combination of phase shifters and splitters and combiners creates modulators to change the amplitude of the in-phase and quadrature components of the signal.
M. The receiver consists of a reference signal, splitters, and balanced photodiodes to combine the signal in-phase and quadrature component with the reference.
N. Signal-to-noise ratio is the signal power divided by the noise power with the system bandwidth. A related parameter, R, is the bit energy over the noise power within the system bandwidth.
O. The signal-to-noise ratio determines the bit error rate for a given communication format.
P. The sources of noise in optical transmission system include noise attributed to the laser, noise attributed to the fiber and amplifiers, and noise attributed to the receiver.
Q. The dominant noise source in transmission systems is quantum noise in the erbium doped fiber amplifier. These amplifiers raise the signal-to-noise ratio of the transmitted signal by 3–6 dB.
R. Shot noise is attributed to absorption of photons being a random, Poisson-distributed process.
S. Johnson (thermal) noise on the receiver is the result of a small, random, thermal current being superimposed on the photocurrent.
T. Independent noise sources combine in quadrature to give a total noise.

12.11 Questions

Q12.1 What is meant by the phase of an optical signal? How would it be directly measured and how is it measured on optical signals?
Q12.2 What is modulated in amplitude shift keying (ASK), phase shift keying (PSK), and quadrature amplitude modulation (QAM) systems?
Q12.3 How are phase shifts in optical signal realized?
Q12.4 Sketch the building blocks of a coherent optical transmitter. What are its essential components?
Q12.5 Sketch the building blocks of a coherent optical receiver. What are its essential components?
Q12.6 How is the mixing required to decode optical phase accomplished?
Q12.7 List the advantages and disadvantages of coherent communication over amplitude shift keying.
Q12.8 What fundamentally limits the transmission capability of an (optical) system?
Q12.9 What are some of the noise sources inherent in optical communications?

12.12 Problems

P12.1 Sketch a phase 'constellation' of the amplitude-shift-keyed transmission system with four levels. Consider how the optical phase matters in an amplitude-modulated system.
P12.2 A quadrature amplitude modulation coherent transmission system is shown in Fig. P12.22 (which include the phase constellation, and some time-domain data showing symbol changes). There are 32 points in the phase constellation phase constellation, and the period is 40 ps.

(a) What is the transmission rate in Gbits/s and the symbol rate in Gbaud?
(b) Assume the arrow in the phase constellation is an amplitude of 1 and a phase of 0 (essentially represents the reference phasor) What is the amplitude and phase of the solid point at 2, 3?
(c) Write the phasor in P12.22 as a time-domain optical signal.

P12.3 In the quadrature amplitude-modulated transmitter pictured in Fig. 12.14, a 90° phase shift is pictured associated with the Q component. How long should that path be to implement a 90° phase shift at 1.55 μm if the waveguide has an effective index of 2.1?
P12.4 Over a certain length, a fiber will attenuate a 1-mW 1.55-μm wavelength average power signal by $\sim$30 dB, down to a 1 μW signal. The signal is then received by a p-i-n receiver.

(a) If the data rate is 10 Gb/s, calculate the number of photons received in a single bit.

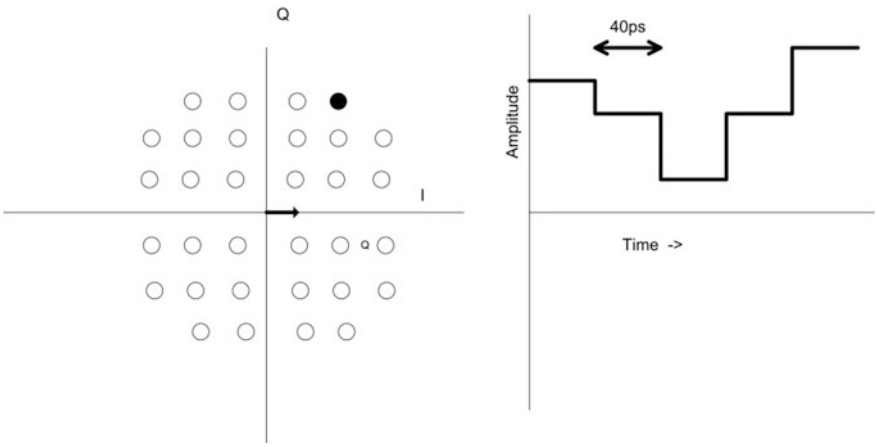


Fig. P12.22 A phasor constellation and sample associated time-domain signal measurements

(b) Calculate the signal/shot noise ratio. (Assume shot noise is the dominant noise in this system.)
(c) From Fig. 12.18, estimate the bit error rate over this link for each of the five formats listed (remember to put it c) in log scale, as the graph is in log scale).

OOK:
BPSK:
8-PSK:
16 QAM:

P12.5 A particular system requires a bit error rate of $<10^{-6}$. For each of the communication formats given, assuming the symbol rate is 10 Gbaud and the noise/bit is 100 nW, what laser power would be needed?

P12.6 A 10 Gbaud/s PAM-4 system has shot noise of 50 nW, thermal (Johnson) noise of 50 nW, and relative intensity noise of 10 nW when received by a photoreceiver after $\sim$100 km of travel. The launch power is 1 mW and the received power is 1 μW.

(a) Find the total noise power and signal-to-noise ratio
(b) Find the energy/bit
(c) Suppose the signal goes through an amplifier with a noise figure of 4 dB, is transmitted another 100 km, and then received. Find the total noise power and signal-to-noise ratio.

Index

A
Absorption, 20–24, 28, 48–50, 52, 69–71, 73–75, 77, 80, 88–91, 93, 95, 101, 103–107, 109, 148, 162, 164, 200, 217, 225, 228, 251, 264, 281, 286, 288, 291, 304–306, 312, 316, 318, 319, 346, 352
Amplifier noise, 346
Arrayed waveguide, 298, 318

B
Back facet phase, 229, 230, 232–235, 239, 240, 251
Balanced photodiode, 334–336, 352
Bandwidth limits, 199, 203, 210, 211
Bernard-Durrafourg condition, 70, 77
Bits, 2, 9, 43, 48, 49, 60, 135, 168, 184, 193, 196, 223, 230, 244, 256–258, 263, 267, 274, 295, 296, 300–303, 317–319, 323, 324, 326–328, 330, 331, 339–345, 349–354
Black body, 2, 12, 13, 15–21, 28, 29, 54, 55
Black body radiation, 12, 13, 19, 29
Bose-Einstein distribution function, 14
Bragg reflector, 152, 220, 222, 223, 226, 244, 250, 275, 288
Built-in voltage, 112, 120–123, 145
Buried heterostructure, 203, 262, 267, 270, 272–275, 280, 288, 290

C
Capacitance, 200, 203, 204, 206–208, 210, 211, 214, 273, 287, 288, 290, 303, 305, 317, 321, 332, 339
Catastrophic Optical Damage (COD), 205, 280, 281
Cavity, 6, 8, 9, 16, 22, 23, 25, 27, 30, 83–88, 91–93, 95, 96, 98–101, 105–110, 151–166, 168, 169, 173, 175, 178–181, 191, 198, 200, 206, 207, 212, 215, 219–222, 224–228, 230, 232, 233, 235, 241, 243–246, 248, 250, 251, 259, 274, 275, 277, 287, 288, 290, 314
Chip, 6, 7, 203, 204, 255, 263, 276–279, 311
Coherent transmission, 312, 317, 331, 338, 339, 351, 353
Coupled mode theory, 226, 240, 241
Coupling, 166, 178, 179, 184, 222, 224, 242, 246, 252, 275, 288, 298, 313, 314, 318, 333, 350
Critical thickness, 31, 39, 40, 42, 50, 51, 53

D
Defects, 40, 91, 93, 94, 205, 263, 264, 269, 272–274, 279–281, 286
Degenerate, 14
Density of states, 8, 9, 13, 15, 17–19, 28, 29, 46, 47, 53–57, 59, 60, 62, 64–68, 70–73, 77–81, 83, 112, 162
Depletion region, 112, 116, 121, 123–126, 128, 129, 131, 137, 142, 145–147
Design, 8, 9, 37, 38, 89, 100, 108, 151, 152, 166, 167, 172, 175, 178, 182, 195, 209, 212, 218, 226, 232, 236, 246, 247, 251, 252, 263–265, 268, 277, 285–288, 290, 315, 320, 331
Design tools, 287
Detuning, 252, 259, 260, 288, 289, 291
D-factor, 198, 201, 206
Differential gain, 47, 108, 187, 191, 192, 199–202, 207, 210–212, 218, 259, 287–289
Differential Phase Shift Keying (DPSK), 326, 327
Diffusion current, 112, 118–120, 124, 128, 129, 145, 149
Diffusion length, 127, 130, 131
Direct bandgap, 43, 44, 46, 50, 51
Dislocation, 40

© Springer Nature Switzerland AG 2020

D. J. Klotzkin, *Introduction to Semiconductor Lasers for Optical Communications*,
https://doi.org/10.1007/978-3-030-24501-6

Dispersion, 4–6, 8, 9, 35, 44–46, 48, 50, 56, 162, 216–218, 240, 256–258, 290, 291, 294–296, 303, 312, 314, 319, 339, 352
Dispersion penalty, 185, 255–258, 289, 291, 303
Distributed feedback laser, 8, 9, 24, 152, 162, 192, 215–218, 222, 224–226, 228, 230, 231, 237–241, 243, 246, 247, 250–252, 256, 258–260, 267, 269–271, 277, 286, 289, 292, 319, 350
Dopant, 35, 113, 115, 116, 118, 120–122, 129, 266
Drift current, 116, 119, 120, 124, 141, 145, 149

E
Effective density of states, 113
Effective index method, 175, 176, 179
Effective mass, 46, 56–59, 63, 65, 77–80, 263
Electrooptic effect, 311, 352
Electrostatic discharge, 285, 286
Erbium Doped Fiber Atmosphere (EDFA), 5, 346
Etalon, 152, 154, 156, 158, 160, 163, 169, 174, 178, 180
External modulation, 8, 184, 293, 296, 303, 312, 317, 318, 351
External quantum efficiency, 98, 104, 105, 107
Eye pattern, 184, 187, 207, 208, 210, 213, 303, 319, 341, 342

F
Fabry-Perot laser, 154
Facet reflectivity, 86, 87, 99, 100, 105, 106, 109, 156, 158, 164, 179, 180, 182, 212, 228, 232
Failures In Time (FITs), 282, 283, 285, 290–292
Far field, 166, 167, 179, 275
Fermi-Dirac distribution function, 28, 66
Fermi level, 9, 66–71, 77, 79, 81, 84, 89, 112–116, 121, 123–126, 132–137, 140, 146–148
Free spectral range, 151, 154, 156–159, 174, 180, 215, 216, 219, 220, 250

G
Gain bandwidth, 160, 161, 179, 215, 216, 219, 220, 222, 250
Gain compression, 191–193, 197, 199–202, 210, 211
Gain coupled, 245
Gain medium, 23, 25–28, 32, 84, 85, 218
Gaussian distribution, 29
Grating fabrication, 249, 267, 270
Group index, 152, 158, 159, 161, 179, 180, 182, 219, 314

H
Hakki-Paoli method, 164

I
Index coupled, 225, 230, 251
Indirect bandgap, 43, 50, 51, 79
Integrated laser modulator, 305, 306
Internal quantum efficiency, 83, 98, 99, 101, 103–107, 189

J
Johnson noise, 348, 349
Joint density of states, 72, 73

K
Kappa, 224, 236
K-factor, 200, 202–205, 208, 210–212, 331, 339

L
Laser bandwidth, 203, 205, 210, 211
Laser bar, 6, 7, 84, 97, 98, 100, 153, 206, 220, 226
Laser linewidth, 258, 312–314
Lateral mode, 165, 166, 179, 180, 219, 275
Lattice constant, 34, 35
Lattice-matched, 34, 39, 42
Linewidth enhancement factor, 313, 314, 318, 320, 321
Lithium niobate, 258, 307–309, 318, 332, 352
Longitudinal mode, 152, 153, 161, 162, 165, 178, 179, 219, 250
Loss coupled, 245

M
Mach-Zehnder modulator, 295, 307, 311, 318, 319
Majority carriers, 115, 124, 125, 127, 129, 135, 138, 141
Matthiessen's rule, 189
Mean Time Before Failure (MTBF), 282–285, 290, 292
Minority carriers, 112, 115, 120, 124–129
Mode index, 159, 160, 165, 174, 179, 182, 200, 235
Modulation, 8, 9, 90, 163, 183–188, 190–192, 196–201, 203–207, 210, 211, 216, 249,

258, 288, 294–296, 299–307, 309–312, 314, 317–320, 323, 328, 333, 339, 340, 351–353
Multiplexing, 217, 294, 296, 298, 299, 317, 329, 330

N
Noise, 5, 9, 205, 206, 217, 294, 302, 312, 314, 323, 326, 329, 335, 336, 339–350, 352–354
Non-radiative lifetime, 212
Non-Return to Zero (NRZ), 300, 301, 319

O
Optical Add Drop Multiplexors (OADM), 299
Optical gain, 8, 53, 54, 65, 70, 71, 76–79, 83–85, 88, 151, 160, 162, 164, 178, 191, 218, 230, 246, 260, 313, 314, 346
Optical loss, 5, 88, 98, 106, 202, 220, 231, 291, 305, 313
Optical overstress, 286

P
Pauli exclusion principle, 14
Phase shift keying, 326, 327, 332, 340, 350, 352, 353
Photon lifetime, 95, 97, 106, 107, 156, 191, 195, 199, 202, 210–213, 287
Photoreceiver, 315, 318, 342, 345, 346, 354
Plasma effect, 310, 311, 313, 318, 332, 352
Polarization division multiplexing, 329, 352
Pulse Amplitude Modulation (PAM), 302, 319

Q
Quadrature Amplitude Modulation (QAM), 328, 333, 352, 353
Quantum-Confined Stark Effect (QCSE), 304, 318
Quantum efficiency, 83, 98, 99, 101, 103–107, 189
Quantum*state*, 14
Quantum well, 6, 8, 9, 18, 31, 34, 39, 40, 42, 43, 50, 51, 53, 55, 59–68, 75, 77, 78, 80, 83, 89, 90, 92, 93, 105, 110, 112, 133, 154, 165, 166, 169, 178, 181, 200, 205, 219, 220, 223, 249, 252, 259, 260, 263, 264, 270, 272, 274, 286–288, 297, 304–306, 318, 319, 351
Quasi-Fermi level, 8, 9, 66–71, 74, 77, 81, 84, 87, 89, 112, 124, 125, 132, 133, 146–148

R
Radiative lifetime, 189, 212
Radiative recombination, 33
Random failure, 280, 290, 292
Reciprocal space, 15, 16, 45
Reduced mass, 72
Reflectivity, 25, 27, 84, 85, 87, 99–101, 105–109, 154, 156, 158, 164, 212, 215, 220–229, 231, 235, 242, 247, 248, 250–253, 277, 286
Relative Intensity Noise (RIN), 206, 207, 345, 349, 354
Reliability, 8, 9, 255, 262, 264, 273, 279–283, 285, 290, 291
Requirements for lasing system, 11, 23
Return-to-Zero (RZ), 300, 318, 319
Ridge waveguide, 6–8, 102, 103, 105, 143, 152, 153, 166, 175, 176, 178, 179, 256, 262, 264, 265, 270–275, 290

S
Schottky junction, 133, 137, 141–143, 324
Shot noise, 345, 346, 349, 352, 354
Side Mode Suppression Ratio (SMSR), 185, 234, 235, 250, 277, 279
Signal-to-noise ratio, 342
Slope asymmetry, 221
Snap test, 286
Space charge region, 116, 118–122
Spatial hole burning, 191, 192, 233, 268, 269
Spectral hole burning, 161, 191, 192, 234
Spontaneous emission, 20–24, 28, 33, 92, 93, 96, 97, 109, 154, 160, 190, 191, 193, 196, 206, 246, 250, 312, 313, 318, 345, 346
Stimulated absorption, 29
Stimulated emission, 11, 12, 20–24, 26–29, 33, 69, 70, 77, 92, 93, 95, 96, 190, 206, 293, 312
Stopband, 244, 246, 247, 250
Strain, 8, 31, 38–43, 50–53, 264, 265, 267
Strain compensation, 42
Submount, 206, 209
Symbols, 9, 78, 198, 244, 295, 300–302, 317–319, 325–329, 331, 332, 335, 339, 340, 342–344, 346, 348, 350–354

T
T_0, 260, 261, 289–291
TE mode, 165, 169, 172, 176, 180
Temperature effects, 258, 259, 289
TM mode, 165, 169, 170, 174, 179–182

Transfer matrix method, 221, 225, 237, 238, 240, 250, 252
Transparency carrier density, 87
Transparency current density, 87, 104–107, 193

U
Unity round trip gain, 85–87, 105, 225, 226

V
V_{π}, 308, 309, 320, 332
Vegard's law, 38
Vertical Cavity Surface Emitting Laser (VCSEL), 108, 109, 212, 219, 220, 274, 275

W
Wafer, 6, 7, 40, 48, 49, 84, 99, 114, 142, 144, 153, 214, 226–228, 231, 232, 235, 247, 249, 251, 263–279, 285, 290, 305, 318
Wafer fabrication, 263, 264, 267, 270, 272, 274
Wavelength Division Multiplexing (WDM), 217, 259, 289, 295–298, 317, 319, 344
Wear out failure, 282, 290
Work function, 133–136, 138–140, 146, 148, 149

Y
Yield, 9, 62, 215, 230, 233, 235, 247, 249, 251, 253, 268, 269, 278, 279, 290